Sources of interference in item and associative recognition memory

Adam F. Osth and Simon Dennis

University of Newcastle, Callaghan, Australia

Address correspondence to:

Adam Osth

E-mail: adamosth@gmail.com

Author Note

This work was presented in partial fulfillment of Adam Osth's dissertation requirement. We would like to thank Per Sederberg, Roger Ratcliff, and Jay Myung for serving on the dissertation committee. Additionally, we would like to thank Larry DeCarlo for generously providing his data and to thank John Wixted, William Hockley, and an anonymous reviewer for helpful suggestions on a previous version of this article.

Abstract

A powerful theoretical framework for exploring recognition memory is the global matching framework, in which a cue's memory strength reflects the similarity of the retrieval cues being matched against the contents of memory simultaneously. Contributions at retrieval can be categorized as matches and mismatches to the item and context cues, including the self match (match on item and context), item noise (match on context, mismatch on item), context noise (match on item, mismatch on context), and background noise (mismatch on item and context). We present a model that directly parameterizes the matches and mismatches to the item and context cues, which enables estimation of the magnitude of each interference contribution (item noise, context noise, and background noise). The model was fit within a hierarchical Bayesian framework to ten recognition memory datasets that employ manipulations of strength, list length, list strength, word frequency, study-test delay, and stimulus class in item and associative recognition. Estimates of the model parameters revealed at most a small contribution of item noise that varies by stimulus class, with virtually no item noise for single words and scenes. Despite the unpopularity of background noise in recognition memory models, background noise estimates dominated at retrieval across nearly all stimulus classes with the exception of high frequency words, which exhibited equivalent levels of context noise and background noise. These parameter estimates suggest that the majority of interference in recognition memory stems from experiences acquired prior to the learning episode.

Sources of interference in item and associative recognition memory

Perhaps the biggest theoretical advance in recognition memory was the application of signal detection theory (SDT) by Egan (1958). SDT recast the role of a participant in a recognition memory experiment as having to decide between whether a presented stimulus is an instance of noise alone (a non-studied stimulus) or signal embedded in noise (a studied stimulus). This is accomplished by comparing the memory strength elicited by a stimulus (which is assumed to be continuously distributed) to a decision criterion on the memory strength axis; stimuli with memory strengths that exceed the decision criterion are judged as having occurred on the study list. Despite the utility of SDT in applications to measurement (Green & Swets, 1966), it is agnostic as to the psychological content of the signal and noise distributions. Specifying the psychological content of the distributions requires process models that describe the encoding and retrieval operations of the memory system along with the content of the stored representations.

A watershed moment in process models of recognition memory came with the *global matching* models of recognition memory. While early theories of recognition memory described the signal and noise distributions as arising from a strength of the stimulus in memory (Wickelgren & Norman, 1966), global matching models, following the encoding specificity principle of Tulving and Thomson (1973), posit that memory strength arises from the similarity between the retrieval cues and the contents of memory. Specifically, the cues are matched against all of the acquired memories in parallel, producing a single memory strength value that indexes the similarity of the cues to the contents of memory (Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989). In the majority of the global matching models, the distance between the signal and noise distributions arises from the match between the target item and its own representation in memory, whereas the variances of the two distributions arise primarily from spurious similarities between the cues and non-target representations stored in memory.

While several models have taken the simplifying assumption that only the memories

from the study list contribute to the retrieval strengths, the frameworks are quite compatible with incorporating contributions from memories learned prior to an experiment. The division between recently acquired and prior memories can be reconciled quite easily by virtue of a *context* representation, which is now featured in the majority of episodic memory models (J. R. Anderson & Bower, 1972; G. D. A. Brown, Preece, & Hulme, 2000; Cox & Shiffrin, 2012; Criss & Shiffrin, 2004; Dennis & Humphreys, 2001; Farrell, 2012; Gillund & Shiffrin, 1984; Howard & Kahana, 2002; Humphreys, Bain, & Pike, 1989; Lehman & Malmberg, 2013; Mensink & Raaijmakers, 1988; Murdock, 1997; Shiffrin & Steyvers, 1997). While there is no universally accepted definition of context, the central assumption among most theorists is that context is what enables retrieval to be focused on a particular episode, namely a study list (Klein, Shiffrin, & Criss, 2007). Learning in contextual models does not merely consist of learning the stimulus, but instead consists of acquiring a binding of the stimulus and a representation of the current context into the contents of memory. At retrieval, the probe cue along with a reinstatement of the study context can be matched against the contents of memory. Under this view, memories of the list items can be distinguished from prior list memories by virtue of the similarity of the stored context representations to the context cues employed at retrieval. Specifically, successful discrimination relies on memories from the study list episode exhibiting more similarity to the context cue, whereas temporally distant memories should be relatively dissimilar to the current context to minimize interference.

The contributions from prior memories and current memories can be conceptualized as matches and mismatches of the stored memories to the item and context cues employed at retrieval, with the magnitudes of each interference contribution determined by the similarities of the matches and mismatches (see Figure 1). Specifically, the locus of successful discrimination is the *self match*, which is a match on stored item and context information to the item and context cues employed at retrieval. Other items from the study list episode match in context information, but mismatch in item information. An

assumption adopted by the earliest global matching models, including Minerva 2 (Hintzman, 1988), the search of associative memory (SAM: Gillund & Shiffrin, 1984) model, the theory of distributed associative memory (TODAM: Murdock, 1982), and the matrix model (Pike, 1984; Humphreys, Bain, & Pike, 1989), is that the studied items of the list episode are the principal source of interference, an idea which has been retroactively referred to as the *item noise* conception of interference.

What was considered a strength of the pure item noise global matching models at the time was their ability to account for the *list length effect* in recognition memory performance, whereby performance decreases as the number of items on a list is increased (Strong, 1912). In global matching models, the spurious similarity between the retrieval cues and a stored memory produces a memory strength value with non-zero variance, and the variances of the resulting distributions are the sums of the variances of the individual matches. The list length effect naturally arises from the early global matching models because only the list items are assumed to be stored in memory and are therefore the principal source of interference. Thus, as more items are added to the contents of memory, the cumulative memory strength is a sum over a larger number of items and the variance in memory strengths for both targets and lures are increased, decreasing discriminability

An unintended consequence of pure item noise models is that the models predict a *list strength effect* in recognition memory performance. A list strength effect occurs when the strengthening of non-target items decreases performance on the target items. Global matching models predicted a list strength effect in recognition memory because item repetitions exhibited the same functional effect as increasing the length of a study list, as each repeated item in memory contributed additional variance to the retrieval process (for a complete description of how each global matching model was unable to predict a null list strength effect, see Shiffrin, Ratcliff, & Clark, 1990). However, Ratcliff et al.'s (1990) investigation found no effect of list strength on recognition memory performance, as the strenghtening a subset of list items did not impair recognition of the non-strengthened

items and strong items did not benefit from being accompanied by weak items on a study list.

The null list strength effect was a strong constraint on the global matching models and various alternatives to the original global matching models were proposed. First, models with a revised encoding process called *differentiation* were proposed that reduce inter-item similarity as the strengths of the study items is increased, allowing for a reduction in item noise with increasing strength such that no detrimental effect of list strength is predicted. Additional items that are added to the contents of memory do not induce differentiation, and thus additional items increase the degree of item noise and a list length effect is predicted. Differentiation models include a modified version of the SAM model (Shiffrin et al., 1990), the retrieving effectively from memory model (REM: Shiffrin & Steyvers, 1997) and the subjective likelihood in memory model (SLiM: McClelland & Chappell, 1998, additional discussion on differentiation models can be found in the General Discussion). However, one of the main motivations behind the differentiation mechanism was to simultaneously predict a null list strength effect while predicting detrimental effects of increasing list length, and more recent evidence suggests that this dissociation may not be present.

Dennis and colleagues (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011) have noted that experiments that manipulate list length contain a number of confounds that may be causing worse performance in conditions with longer lists for reasons unrelated to interference among the list items. For instance, the retention intervals are shorter for short lists than long lists if testing immediately follows the end of the study list, reducing performance for items on the long list. When this confound and others are controlled, all experiments conducted by Dennis and colleagues that employed words as stimuli have found no effect of list length on discriminability. Two other modifications to the global matching framework can allow for predicting null effects of list length and list strength on recognition memory performance.

Dennis and Humphreys (2001) argued that interference arises not from the list items (as assumed by pure item noise models) but from memories of the list items acquired prior to the experiment. Specifically, the retrieval cues are not just matched against representations from the list episode, but are matched against all of the contexts in which the items were experienced to evaluate whether the context of the study list is included in the set of all stored context representations. This conception of interference has been referred to as *context noise*, because it is the past contexts in which the list items have been experienced (context mismatch) that generate interference in recognition memory (an idea which originated from J. R. Anderson & Bower, 1972). Dennis and Humphreys (2001) introduced the bind-cue-decide model of episodic memory (BCDMEM), in which context noise was the sole source of interference in the model. Item representations in the model do not overlap with each other, meaning that the item mismatch penalty is zero and no effects of item noise, namely list length and list strength, are predicted. Dennis and Humphreys (2001) also demonstrated that the word frequency effect, in which words of low natural language frequency are better recognized than words of high natural language frequency (Glanzer & Bowles, 1976; Glanzer & Adams, 1985; Shepard, 1967), follows quite naturally from the concept of context noise: items that have been more frequently experienced have more associations to prior contexts, and thus there is more ambiguity as to whether or not they were seen in a given context. This is analogous to the manner in which item noise models predict a list length effect in the sense that more stored representations produce additional interference at retrieval.

Another solution to the list strength effect came from Murdock and Kahana (Murdock, 1997; Murdock & Kahana, 1993a, 1993b), who posited that global matching models should include a large number of memories that mismatch in both item *and* context information. This conception of interference has since been referred to as *background noise* by Osth, Dennis, and Kinnell (2014). Murdock and Kahana (1993a) argued that if the contribution of background noise is large relative to the item noise from the list items, then

increases in interference that come from a list strength manipulation will produce only a negligible increase in the variances of the memory strength distributions, thus producing an approximate null list strength effect. Osth et al. (2014) revisited the idea of background noise for explaining the small item noise effects found for novel non-linguistic stimuli.

Kinnell and Dennis (2012) and Osth et al. (2014) conducted experiments using images of non-famous faces, random scenes, and generated fractals that are unlikely to have been ever witnessed by the participants prior to the experiment. Given the novelty of the stimuli, they cannot suffer from context noise as they are unlikely to have been seen in prior contexts. Unlike words, small effects of list length (Kinnell & Dennis, 2012) and list strength (Osth et al., 2014) were found for select non-linguistic stimuli employing the list length controls advocated by Dennis and colleagues. While detrimental effects of list length and list strength are not predicted by context noise models, they were much smaller than what would be predicted by pure item noise models. Osth et al. (2014) posited that novel non-linguistic stimuli might suffer from larger item noise than words, possibly due to exhibiting a higher degree of within-class inter-item similarity than words, but the effects of list length and list strength are somewhat mitigated by the additional influence of background noise at retrieval. Background noise has received relatively little attention in the recognition memory literature compared to the discussions of the influence of item and context noise. A diagram depicting the three sources of interference (item noise, context noise, and background noise) can be seen in Figure 1. The present investigation is focused on simultaneously measuring all three sources of interference by fitting a global matching model to a large number of recognition memory datasets.

Measuring Interference Contributions Within a Single Global Matching Model

A number of investigations have compared the relative merits of the item noise and context noise approaches to recognition memory using experimental data (Cho & Neely, 2013; Criss, Malmberg, & Shiffrin, 2011; Dennis & Chapman, 2010; Dennis et al., 2008;

Kinnell & Dennis, 2011), with interpretations favoring either the item noise or context noise accounts, while there was little discussion of the role of background noise in any of these investigations. More recently, Turner, Dennis, and Van Zandt (2013) compared the REM (the original pure item noise version) and BCDMEM models in their ability to account for data from the list length paradigm (specifically the datasets of Dennis et al., 2008 and Kinnell & Dennis, 2012) using hierarchical Bayesian methods. BCDMEM consistently exhibited lower values of the deviance information criterion (DIC), a Bayesian model selection measure that measures goodness of fit relative to the degree of model complexity. Turner, Dennis, and Van Zandt (2013) attributed the superior performance of the BCDMEM model to the fact that the null list length effect in the data is a compulsory prediction of the BCDMEM model, whereas REM exhibits flexibility in its predicted magnitude of the list length effect. The higher DIC value assigned to the REM model may indicate that this flexibility is an unwarranted complexity of the model, and the authors attributed this strength to a parsimony of the context noise account.

However, as noted by Criss and Shiffrin (2004), item noise and context noise are not mutually exclusive; it is completely plausible for a memory system to suffer from both item noise and context noise at retrieval. Not only does there remain the underexplored interference contribution of background noise, but there is also the possibility that the magnitude of each interference contribution depends on the stimulus class being employed. In this article, we present the results of fitting a global matching model to a large number of recognition memory datasets and measuring the respective contributions of item noise, context noise, and background noise to the total interference contribution at retrieval. The model is a variant of the tensor model of Humphreys, Bain, and Pike (1989), in which memory is a composite of three-way bindings between two items and the experimental context.

We deviate from the approach used in the original tensor model and several other vector based models by avoiding specification of the vectors. That is, the standard

approach is to generate item and context vectors from a sampling distribution with a finite number of elements, and model predictions are derived by calculating the dot products between vectors to index the strength of the matches. This approach requires one to commit to parameters such as the number of elements in the vector, which typically do not have a direct psychological interpretation. In our approach, we avoid specifying the vectors and instead parameterize the similarities between the item and context vectors. As we will later demonstrate in the paper, the parameters of the model can be used to analytically calculate the magnitude of the item noise, context noise, and background noise contributions. The datasets included in the fit include manipulations of all of the variables that are required to constrain the parameters of the model, such as strength, word frequency, list length, list strength, and study-test delay.

Furthermore, a critical limitation of the BCDMEM model was its inability to account for stimuli other than single words (which have the necessary background experience to suffer from context noise) and that it lacks a mechanism for inter-item binding, which prevents extension to associative memory tasks such as associative recognition. As we have mentioned previously, experiments with novel non-linguistic stimulus classes uncovered small detrimental effects of list length and list strength, which are consistent with item noise models but inconsistent with context noise models (Kinnell & Dennis, 2012; Osth et al., 2014). However, as noted by Osth et al. (2014), the detrimental effects of list length and list strength were quite small in magnitude as compared to what would be expected from a pure item noise model, and they suggested that both item noise and background noise are relevant to understanding recognition memory performance for nonlinguistic stimuli. The model we are presenting is capable of addressing these sources of interference and we have included the experiments conducted in these two papers to compare the interference contributions across the different stimulus classes. While one might be concerned that including all interference sources produces a more flexible model, the fact that nonlinguistic stimuli may meaningfully differ in their susceptibility to these different

interference sources justifies a comprehensive model.

Additionally, we have included experiments conducted using the associative recognition task, in which participants study a list of pairs (such as A-B, C-D, E-F, etc.) and are asked to discriminate between studied pairs (such as A-B, referred to as *intact* pairs) and studied words presented in a novel arrangement (such as C-F, referred to as *rearranged* pairs). There has been relatively little discussion as to the sources of interference in the associative recognition task in the literature. We include the results of two experiments, one that manipulated list length (Kinnell & Dennis, 2012) and one that manipulated list strength (Osth & Dennis, 2014) to measure the sources of interference in associative recognition.

The outline for the remainder of the paper is as follows. First, we describe our variant of the Humphreys, Bain, and Pike (1989) model and how it calculates the three sources of interference that have been postulated to affect recognition memory. We also discuss a necessary addition to the model to address the mirror effects present in our data, namely the log likelihood ratio transformation of memory strengths by Glanzer, Hilford, and Maloney (2009). Next, we give a summary of the ten datasets that were used in the model fitting along with a description of how the parameters used in the fitting matched the experimental manipulations. We then describe how the models were fit using hierarchical Bayesian methods to get simultaneous estimates of both subject and group level parameters. We then present the results of the model fitting procedure along with analyses of the resulting group and subject level parameters to compare the respective contributions of the sources of interference.

The Model

We follow the tradition of several memory models and represent both items and contexts as vectors of features. To simplify description as much as possible, we define items as single stimuli that are presented to the participant and the context as a representation

that defines the list episode. We follow several other episodic memory models in our assumption that item features and context features are independent of each other (G. D. A. Brown et al., 2000; Criss & Shiffrin, 2004; Dennis & Humphreys, 2001; Humphreys, Bain, & Pike, 1989; Mensink & Raaijmakers, 1988; Murdock, 1997; Shiffrin & Steyvers, 1997)[1].

Bindings between items and contexts are represented as outer products of the constituent item and context vectors. Each element of an outer product is a multiplication of elements in the constituent vectors. A similar way to represent bindings is the convolution operation, in which the diagonals of the outer product matrix are summed together, reducing the outer product to a vector (e.g.: Murdock, 1982; Eich, 1982; Jones & Mewhort, 2007). Both the outer product and convolution bindings are similar in that they are both conjunctive representations of their participating constituents, rather than linkages between nodes in an associative network. Conjunctive representations are associations that are represented much in the same way as individual items are, but bear little similarity to their constituent item vectors. We have chosen to use the outer product over the convolution to represent binding because it is more analytically tractable and simpler, as the additional summation in the convolution introduces noise into the binding (Pike, 1984).

Evidence supporting conjunctive representations comes from a study by Dosher and Rosedale (1989) in which participants studied triplets of items and were tested on pairs from the triplets in an associative recognition task. Successful priming was only found when the entire triplet was completed by the prime, such as if a triplet ABC was studied and item A preceded the pair BC. Dosher and Rosedale (1989) found no evidence for priming on partial matches, such as item A preceding a BF pair trial. Similarly, Hockley

[1]Another possible assumption that is employed by the temporal context model (Howard & Kahana, 2002) is that context features are the previously encountered items, causing a high correlation between item features and context features. However, models that employ this assumption of context have yet to be applied comprehensively to data from recognition memory paradigms.

and Cristi (1996b) conducted a judgment-of-frequency (JOF) task in which both items and pairs were studied and participants made JOFs on both studied items and pairs. Some of the items that were presented alone were constituents of the pairs, e.g.: item A presented alone, and also as part of pair A-B. Hockley and Cristi (1996b) found that the frequency of A had no influence on the judged frequency of A-B, as if the A representations and A-B representations did not overlap with each other.

When a list of items is studied, the model stores outer products of the item and the list context on each trial. These outer products are then summed together to produce an occurrence matrix M_i:

$$M_i = \sum_a r_{item} C_s \otimes I_a \tag{1}$$

where C denotes a context vector, I denotes an item vector, and the subscript s refers to the fact that the context vector represents the study episode. To account for variation in strength of learning, due either to different rates of presentation, different numbers of presentations, or differences among participants in their ability to encode the material, we use a scalar r_{item} that is applied to the outer products as a learning rate parameter.

When a list of pairs are studied for an associative recognition task, the model stores three-way outer products between the two items in the pair and the list context as a mode three tensor product. These tensor products are then summed together to produce the co-occurrence tensor M_c:

$$M_o = \sum_{a,b} r_{assoc} C_s \otimes I_a \otimes I_b \tag{2}$$

where r_{assoc} is the learning rate for associative information. We would like to emphasize that we employ a separate tensor representation for associative recognition purely for mathematical convenience and are not committed to the idea that the occurrence matrix and co-occurrence tensor reflect different neurological substrates or stores. We also make the simplifying assumption that when participants are studying word pairs in an

associative recognition task the only inter-item associations that are formed are among the pair members (a similar assumption was made by Gillund & Shiffrin, 1984).

We allotted a separate learning rate parameter for associative information (r_{assoc}) based on the finding that encoding manipulations produce different effects on item and associative recognition. For instance, Hockley and Cristi (1996a) found that deep encoding manipulations that emphasize item information enhance item recognition but punish associative recognition, whereas deep encoding manipulations that emphasize associative information enhance both item and associative recognition. Thus, our model allows for the possibility that encoding strength can be strong in both item and associative recognition, weak in one task but not the other, etc.

Memory strength is computed by combining the cues available at retrieval into an outer product and matching it against the appropriate memory store. In the case of item recognition, this involves construcing an outer product of the probe cue and the context cue employed at retrieval. This matrix is matched against the occurrence matrix M_i:

$$s = (C'_s \otimes I'_a).M_i \tag{3}$$

where s is a scalar that represents the memory strength generated from the global match of the cues against the contents of memory. The dashes on the context and item vectors are used to indicate that the item vector representing the probe and the context vector representing the context at the time of test may not perfectly resemble the vectors that were used at the time of study. A cue for a target item may not resemble the vector that was originally stored due to variation in perceptual processing of the stimulus (McClelland & Chappell, 1998). The context cue employed at test may not resemble the study context because it is either an imperfect reinstatement of the study context (G. D. A. Brown et al., 2000; Dennis & Humphreys, 2001) or a context that has drifted from the original stored study context as a consequence of the events that have intervened between study and test (Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Murdock, 1997). For the present

purposes, we are agnostic as to whether the context cue employed at test has merely drifted from its original representation or was actively reinstated at test. However, it should also be mentioned that the question of how a context representation could be reinstated is an unsolved problem in contextual models of episodic memory.

The equation is the same for associative recognition, except it involves combining both item cues with the context vector into a tensor representation and matching it against the co-occurrence tensor M_c. The equation is as follows:

$$s = (C'_s \otimes I'_a \otimes I'_b).M_c \tag{4}$$

It is common at this stage for the vectors to be generated from sampling distributions with a finite number of elements. The number of elements in a vector can be considered a parameter of the model despite the fact that it contains no obvious psychological interpretation. When this parameter is free to vary in a model fit, the model parameters are no longer able to be identified in models such as REM (Montenegro, Myung, & Pitt, 2011) and BCDMEM (Myung, Montenegro, & Pitt, 2007). As a consequence, it is common practice for the vector size parameter in a model to be fixed to an arbitrary value.

We instead use an approximate analytic solution that specifies the similarities between the vectors without specifying the content of the vectors themselves. Such an approach is convenient as it allows the different sources of interference in retrieval to be parameterized as matches and mismatches among the context and item vectors. The analytic solution is obtained by decomposing the retrieval equation into each of the component matches in a manner similar to that used by Humphreys, Pike, et al. (1989) in their analyses of the global matching models. An advantage of the analytic solution is that the explicit likelihood function of the model's predictions allows the model to be efficiently fit using Markov chain Monte Carlo (MCMC) methods.

For the case of item recognition, Equation 3 can be rewritten by decomposing the occurrence matrix M_i into matches among all of the stored memories. If the probe item is

a target item, the equation is as follows:

$$
\begin{aligned}
s = (C'_s \otimes I'_t).[r_{item}(C_s \otimes I_t) & \quad \text{Self Match} \\
+ \sum_{i \in L, i \neq t} r_{item}(C_s \otimes I_i) & \quad \text{Item Noise} \\
+ \sum_{u \in P, u \neq s} (C_u \otimes I_t) & \quad \text{Context Noise} \\
+ \sum_{u \in P, u \neq s, z \notin L} (C_u \otimes I_z)] & \quad \text{Background Noise}
\end{aligned}
\tag{5}
$$

The first term in the right column is the original studied item in the study list context. The match between this matrix and the matrix cue can be referred to as the self match (match to both item and context cues) which is not present in the matching equation for a lure. The self match determines the difference in the means between the signal and noise distributions.

The second term in the right column is all of the study list items that are not the target item. The L subscript refers to the set of all of the list items. Similarity between the cue item I'_t and the stored list items produces item noise. As was previously mentioned, the majority of the original global matching models tended to only consider self matches and item noise and never considered the role of pre-experimental interference. Nonetheless, interference from pre-experimentally stored memories could be expected to play a role in memory retrieval, and we consider their possible matches below.

The third term in the right column is the match of the probe item to all of its pre-experimentally stored representations. The u subscript of the context vector denotes that these stored contexts are different from the study list context and the U subscript in the sum refers to the set of all contexts over a lifetime that are not the study list context. Similarity among the reinstated context cue C'_s and the pre-experimental contexts produces context noise at retrieval. The BCDMEM model can be considered an example of a global matching model that only considers the self matches and context noise at retrieval.

The fourth term in the right column is the match of the probe item to everything else that has been stored in memory. That is, all memories that mismatch in both item and

context information are contained in this term. If these memories overlap with the matrix cue they would produce interference that we refer to as *background noise*. As we have mentioned previously, this term does not contribute in most memory models, with the exceptions of the TODAM (Murdock & Kahana, 1993a, 1993b) and TODAM2 (Murdock, 1997) models along with a variant of the SAM model presented by Gronlund and Elam (1994). Equation 5 can be rewritten as the match between the test cues and the stored vectors in memory:

$$\begin{aligned} s =& r_{item}(C'_s.C_s)(I'_t.I_t)+ && \text{Self Match} \\ & \sum_{i\in L, i\neq t} r_{item}(C'_s.C_s)(I'_t.I_i)+ && \text{Item Noise} \\ & \sum_{u\in P, u\neq s} (C'_s.C_u)(I'_t.I_t)+ && \text{Context Noise} \\ & \sum_{u\in P, u\neq s, z\notin L} (C'_s.C_u)(I'_t.I_i) && \text{Background Noise} \end{aligned} \tag{6}$$

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:

$$\begin{aligned} C'_s.C_s &\sim Normal(\mu_{ss}, \sigma^2_{ss}) && \text{Context Match} \\ C'_s.C_u &\sim Normal(\mu_{su}, \sigma^2_{su}) && \text{Context Mismatch} \\ I'_t.I_t &\sim Normal(\mu_{tt}, \sigma^2_{tt}) && \text{Item Match} \\ I'_t.I_i &\sim Normal(\mu_{ti}, \sigma^2_{ti}) && \text{Item Mismatch} \end{aligned} \tag{7}$$

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 appplication by allowing separate specification of the mean and variance parameters. As we will discuss below, this is necessary to avoid covariances.

The matches in Equation 7 include the match between the test context and the context of study (context match, indexed by subscript *ss*), the match between the test context and contexts prior to the study list (context mismatch, subscript *su*), the match between the probe cue and its own stored item representation (item match, subscript *tt*), and the match between the probe cue and other items stored in memory (item mismatch, subscript *ti*).

The distributions of the matches and mismatches from Equation 7 are substituted into the terms for Equation 6 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:

$$\mu_{old} = r_{item}\mu_{ss}\mu_{tt} + r_{item}(l-1)\mu_{ss}\mu_{ti} + m\mu_{su}\mu_{tt} \quad (8)$$

$$\mu_{new} = r_{item}l\mu_{ss}\mu_{ti} + m\mu_{su}\mu_{tt} \quad (9)$$

$$\begin{aligned}\sigma^2_{old} =& r^2_{item}(\mu^2_{ss}\sigma^2_{tt} + \mu^2_{tt}\sigma^2_{ss} + \sigma^2_{ss}\sigma^2_{tt})+ && \text{Self Match} \\ & r^2_{item}(l-1)(\mu^2_{ss}\sigma^2_{ti} + \mu^2_{ti}\sigma^2_{ss} + \sigma^2_{ss}\sigma^2_{ti})+ && \text{Item Noise} \\ & m(\mu^2_{su}\sigma^2_{tt} + \mu^2_{tt}\sigma^2_{su} + \sigma^2_{su}\sigma^2_{tt})+ && \text{Context Noise} \\ & n(\mu^2_{su}\sigma^2_{ti} + \mu^2_{ti}\sigma^2_{su} + \sigma^2_{su}\sigma^2_{ti}) && \text{Background Noise}\end{aligned} \tag{10}$$

$$\begin{aligned}\sigma^2_{new} =& r^2_{item}l(\mu^2_{ss}\sigma^2_{ti} + \mu^2_{ti}\sigma^2_{ss} + \sigma^2_{ss}\sigma^2_{ti})+ && \text{Item Noise} \\ & m(\mu^2_{su}\sigma^2_{tt} + \mu^2_{tt}\sigma^2_{su} + \sigma^2_{su}\sigma^2_{tt})+ && \text{Context Noise} \\ & n(\mu^2_{su}\sigma^2_{ti} + \mu^2_{ti}\sigma^2_{su} + \sigma^2_{su}\sigma^2_{ti}) && \text{Background Noise}\end{aligned} \tag{11}$$

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 10 can be viewed as the contributions of the self match, item noise, context noise, and background noise. Equations 10 and 11 are identical with the exception of the self match variance term which is only in Equation 10 and the fact that item noise is scaled by $l-1$ in Equation 11 instead of l.

These equations also reveal the various effects of item noise, context noise, and background noise: each of them contribute additional variance to both the old and new distributions. What disentangles these interference sources is the selective influence of experimental manipulations. Increases in the number of list memories l and increases in the learning rate r increase the item noise, increases in the prior occurrences of the cue m increase the context noise, and increases in the number of other stored memories n increases the background noise. All of these manipulations increase the total variance in a linear fashion.

Some simplicitations are made in order to reduce the number of parameters and avoid covariances. Specifically, in Equation 6, one can see that there are multiple item matches

and context matches across the different interference terms. To avoid covariances between these matches, we fix the parameters μ_{ti} and μ_{su} to zero, a simplification which also has the effect of fixing the mean of the lure distribution at zero. There is precedent for such an approach, as both the Minerva 2 (Hintzman, 1988) and TODAM (Murdock, 1982) models have a lure distribution that is fixed at zero by usage of zero-centered vectors, which ensures that the expected match between any two vectors representing different items is zero.

In addition, we are more interested in the variance contribution in the context noise and background noise term than we are in identifying the number of stored memories. For that reason, we ignore the m term and instead allocate separate context mismatch variability parameters to high and low frequency items to reflect the varying degrees of context noise. We denote the combined influence of m and σ_{su}^2 as parameter ρ. Additionally, we eliminate the entire background noise term and instead substitute a separate variance parameter to reflect its contribution, which we denote as β. The simplified equations are as follows:

$$\begin{aligned} \mu_{old} =& r_{item}\mu_{ss}\mu_{tt} \\ \mu_{new} =& 0 \end{aligned} \tag{12}$$

$$\begin{aligned} \sigma_{old}^2 =& r_{item}^2(\mu_{ss}^2\sigma_{tt}^2 + \mu_{tt}^2\sigma_{ss}^2 + \sigma_{ss}^2\sigma_{tt}^2)+ && \text{Self Match} \\ & r_{item}^2(l-1)(\mu_{ss}^2\sigma_{ti}^2 + \sigma_{ss}^2\sigma_{ti}^2)+ && \text{Item Noise} \\ & (\mu_{tt}^2\rho + \rho\sigma_{tt}^2)+ && \text{Context Noise} \\ & \beta && \text{Background Noise} \end{aligned} \tag{13}$$

$$\begin{aligned} \sigma^2_{new} =& r^2_{item} l(\mu^2_{ss}\sigma^2_{ti} + \sigma^2_{ss}\sigma^2_{ti}) + && \text{Item Noise} \\ & (\mu^2_{tt}\rho + \rho\sigma^2_{tt}) + && \text{Context Noise} \\ & \beta && \text{Background Noise} \end{aligned} \tag{14}$$

Each interference term in Equations 13 and 14 arises from combinations of the matches and mismatches of context and item information. The mean of the target distribution is a multiplication of the learning rate r_{item}, the mean of the item match μ_{tt}, and the mean of the context match μ_{ss}. In our fits of the model to data, we fixed the mean of the item match μ_{tt} to one for simplicity[2]. We vary the mean context match μ_{ss} across conditions that vary in retention interval to reflect the loss of study context information from contextual drift or imperfect reinstatement of the study context. For conditions where testing is either immediate or follows shortly after the study list, we fix the value of μ_{ss} at one.

All mismatch parameters contribute to the variances of the distributions rather than the means. The self match variability is a function of the mean and variances of the item and context matches (as we will see below, appropriate choices of these parameter values can instill higher variance in the target distribution than the lure distribution). The item noise term is a function of the number of list items multiplied by the variability in the item mismatch σ^2_{ti}, which is scaled by both of the context match parameters. The context noise term is a function of the variability in the context mismatch ρ scaled by both of the item match parameters.

Both the self match and item noise terms are scaled by the learning rate. Thus, as the encoding strength is increased, both the self match variability and the item noise are increased. The increase in item noise is the locus of the list strength effect. In the list

[2]It would be plausible for the mean of the item match to vary across conditions that vary in stimulus strength. It is extremely plausible that the different stimulus classes in our investigation vary in item strength, but was simpler to assume that all of the differences arose from differences in learnability.

strength datasets we will be considering in this article, two list conditions are used: one where all items are presented once (the *pure weak* condition) and one where half the items are presented four times and half the items are presented once (the *mixed* condition). Specification of mixed list interference requires learning rates for weak and strong items along with separate item noise terms for each learning rate that are added together. A list strength effect would be present if performance on the once presented weak items is worse in the mixed list relative to the pure weak list, and arises if the item noise from the strong items is sufficiently larger than the item noise from the weak items. As we will later demonstrate, a null list strength effect can be predicted if either a.) item noise is sufficiently low from a low value of the item mismatch parameter σ_{ti}^2 or b.) the ratio of background noise to item noise is sufficiently high such that the strong item interference presents only a negligible addition to the total interference.

To avoid redundancy in the text, derivations of the distributions for associative recognition can be found in Appendix A. The same distributions for the item and context matches and mismatches are used. The mean and variances of the resulting memory strength distributions for intact and rearranged pairs are as follows:

$$\begin{aligned} \mu_{int.} =& r_{assoc}\mu_{ss}\mu_{tt}^2 \\ \mu_{rearr.} =& 0 \end{aligned} \tag{15}$$

$$\begin{aligned} \sigma_{int.}^2 =& r_{assoc}^2(2\mu_{ss}^2\mu_{tt}^2\sigma_{tt}^2 + \mu_{ss}^2\sigma_{tt}^4 + \sigma_{ss}^2\mu_{tt}^4 + 2\sigma_{ss}^2\mu_{tt}^2\sigma_{tt}^2 + \sigma_{ss}^2\sigma_{tt}^4) + && \text{Self Match} \\ & r_{assoc}^2(l-1)(\mu_{ss}^2\sigma_{ti}^4 + \sigma_{ss}^2\sigma_{ti}^4) && \text{Item Noise} \\ & \beta_{assoc} && \text{Background Noise} \end{aligned} \tag{16}$$

$$\begin{aligned}\sigma^2_{rearr.} =& 2r^2_{assoc}(2\mu^2_{ss}\mu^2_{tt}\sigma^2_{ti} + \mu^2_{ss}\mu^2_{tt}\sigma^2_{ti} + \sigma^2_{ss}\mu^2_{tt}\sigma^2_{ti} + \sigma^2_{ss}\sigma^2_{tt}\sigma^2_{ti}) + && \text{Partial Match} \\ & r^2_{assoc}(l-2)(\mu^2_{ss}\sigma^4_{ti} + \sigma^2_{ss}\sigma^4_{ti}) + && \text{Item Noise} \\ & \beta_{assoc} && \text{Background Noise}\end{aligned} \tag{17}$$

The interference terms are very similar to the derivations for item recognition with a few exceptions. The first is that there is no context noise term; this is because the associative recognition task involves random pairings of unrelated words and thus the probability of having seen a given pair multiple times prior to the experiment is negligible (this assumption would not apply if previously associated pairs are employed). Second, there is an additional partial match term for the rearranged pairs in Equation 17. This reflects the fact that for a rearranged pair A-D, there are two partially matching pairs in memory: A-B and C-D. Additionally, given that there are likely to be many more combinations of item-item-context bindings than single item-context bindings, a separate background noise term was allotted for the co-occurrence tensor (β_{assoc}). All of the model parameters along with their psychological interpretations and experimental manipulations that change them can be found in Table 1. Next, we describe the likelihood ratio transformation of memory strengths that is necessary to capture the full range of mirror effects seen in recognition memory data.

Table 1

Description of each of the model's parameters, including their boundaries and which conditions they change.

Param	Bounds	Description
r_{item}; r_{assoc}	0 : 1	Learning rates for items and associations. Increase with study time or repetitions.

Continued on next page

Table 1 - continued from previous page

Param	Bounds	Description
μ_{ss}	$0:1$	Context match mean: matching strength of test context cue to stored context. Contributes to the mean of the target distribution. Decreases with study-test delay.
σ^2_{tt}	$0:\infty$	Item match variability: Variability of the match of the item cue to the stored item. Increases the variability of the target distribution relative to the lure distribution.
σ^2_{ss}	$0:\infty$	Context match variability: Variability of the match of the context cue to the stored context. Increases the variability of the target distribution relative to the lure distribution.
ρ	$0:\infty$	Context mismatch variability: Contributes to the amount of context noise in the model. Expected to vary with word frequency and is zero for items not seen in prior contexts.
σ^2_{ti}	$0:\infty$	Item mismatch variability: Contributes to the amount of item noise in the model. Magnifies effects of list length and list strength manipulations. Varies by stimulus class.
β_{item}; β_{assoc}	$0:\infty$	Background noise for items and associations. Obscures effects of list length/list strength on performance. Varies by stimulus class.
Φ	$-\infty:\infty$	Response criterion. 0 represents an unbiased criterion.

The Context Mismatch Parameter, the Word Frequency Effect, and the Likelihood Ratio Transformation of Memory Strengths

A strong constraint on models of recognition memory is the *word frequency mirror effect*, in which low frequency (LF) words have higher hit rates and lower false alarm rates than high frequency (HF) words[3]. The mirror effect was described as a challenge to simple strength models of recognition memory (Glanzer & Adams, 1985, 1990). As we will demonstrate, this is partially true. The basic pattern of the mirror effect can be achieved using the memory strength computation in our model, but there is evidence from two alternative forced choice (2AFC) testing that suggests a mirror ordering in the means of the distributions, which requires a likelihood ratio transformation of memory strengths.

The locus of the word frequency effect in the model is the context mismatch variability parameter ρ, which primarily contributes to the context noise term in Equations 13 and 14 for item recognition. From inspection of the equations, one can see that context noise is produced by a multiplication of the item match parameters μ_{tt} and σ_{tt}^2 along with the context mismatch variability parameter ρ. Critically, when context mismatch variability is zero, there is no context noise. This reflects the idea that there is no similarity between the current context and the previously stored contexts: the current context cue C_s' is perfectly able to isolate the list items from pre-experimentally stored memories. This is an implicit assumption in the early global matching models that assumed that only the list items contributed to interference at retrieval.

When context mismatch variability is greater than zero, greater interference arises from items that are more frequently represented in memory. As a consequence, high frequency words suffer more interference than low frequency words. Model predictions with different values of the context mismatch parameter ρ can be seen in the middle panel of Figure 2. It is interesting to note that the model is able to predict a mirror effect of word

[3]A mirror effect refers to any manipulation that exerts opposite effects on the hit rates and false alarm rates (Glanzer & Adams, 1985)

frequency: as frequency increases, hit rates decrease and false alarm rates increase. This is because context noise, like item noise, is a variance term for both target and lure items. For a fixed decision criterion, an increase in variance of both distributions will cause a mirror effect.

Glanzer and Bowles (1976) conducted a thorough test of the locus of the word frequency effect using 2AFC tests. In a 2AFC tests, it is assumed that a response criterion for a stimulus is not used. Instead, the choice is selected that is furthest on the decision axis (T. D. Wickens, 2002). Glanzer and Bowles (1976) manipulated the composition of the choices on 2AFC tests, using all possible combinations of old and new items such as LF-old and LF-new trials (LO-LN), LF-old and HF-new trials (LO-HN), HF-old and LF-new trials (HO-LN), and HF-old and HF-new trials (HO-HN). The mirror effect was obtained in all cases, and the ordering of the probability of correct choice was as follows: LO-LN $>$ LO-HN $\approx$ HO-LN $>$ HO-HN. While the occurrence of the mirror effect in 2AFC testing is challenging to criterion shift accounts of the mirror effect (e.g.: Gillund & Shiffrin, 1984), inspection of the top right panel of Figure 2 reveals that the variance account in our model is capable of addressing this pattern (LO-LN $>$ LO-HN $\approx$ HO-LN $>$ HO-HN).

However, there two more trial types that the variance account is unable to address. Both of these trial types can be considered *null comparisons* because they are not valid trials with one correct choice and one incorrect choice. In both types of trials, one word is LF and the other is HF, but in one there are two targets (LO-HO) and in another there are two lures (HN-LN). Surprisingly, in the target trials, the LF word is chosen more often (p(LO, HO) $>$ .5) but in the lure trials, the HF word is chosen more often (p(HN, LN) $>$ .5). One can see that in Figure 2, the variances account fails to produce choice probabilities that are greater than .5 for the null comparison trials. Glanzer and Bowles (1976) noted that what is necessary is a mirror ordering arrangement of the means of the signal and noise distributions, such as LN $<$ HN $<$ HO $<$ LO. Mirror arrangements of the underlying distributions can be produced using a log likelihood ratio transformation of the

memory strengths.

The Log Likelihood Ratio Transformation. Several current models of recognition memory employ likelihood ratios as the basis of a recognition memory decision to produce the mirror effect (Dennis & Humphreys, 2001; Glanzer & Adams, 1990; Glanzer, Adams, Iverson, & Kim, 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In log likelihood ratio models, decisions are not made on the basis of memory strength, but are instead made on the basis of a ratio of the densities of the signal and noise distributions. To understand how the mirror effect is derived from a likelihood ratio, take a point on the x axis of the signal and noise distributions in the top left panel of Figure 2. Compare the relative heights of the targets and lures of the HF distributions at that point: that ratio is the likelihood ratio for the stimulus that elicited that value of memory strength. On the same point on the x axis, consider the LF distributions. Note that in the case of target items, the density of the target distribution greatly exceeds the density of the lure distribution due to the lower overlap of the distributions, producing higher likelihood ratios. LF words have lower overlap among the target and lure distributions than HF words, producing higher likelihood ratios and increasing their hit rate. The opposite is the case for lures, in that there is greater lure-to-target density for LF words, producing lower likelihood ratios and a lower false alarm rate relative to HF words. A psychological interpretation of the likelihood ratio transformation is that memory strength is not considered alone, but is instead considered along with the knowledge about the memorability of the stimlus (similar to the account proposed by J. Brown, Lewis, & Monk, 1977). As we will see, the expected memorabilities need not correspond perfectly to the actual memory strength distributions.

The mirror effect has been demonstrated to be a regularity of the likelihood ratio transformation (Glanzer et al., 1993, 2009). There are other advantages to the likelihood ratio transformation, such as the concentering of the likelihood ratio distributions in response to manipulations that decrease performance (Glanzer, Adams, & Iverson, 1991;

Hilford, Glanzer, & Kim, 1997; Kim & Glanzer, 1993, 1995), shorter zROC lengths for conditions of stronger performance (Stretch & Wixted, 1998a; Glanzer et al., 2009), as well as higher variances for distributions that are further from the criterion (as measured by old-old and new-new zROC slopes: DeCarlo, 2007; Glanzer et al., 2009). Additional discussion on data that have supported or challenged likelihood ratio models can be found in the General Discussion.

We have employed analytic solutions for the log likelihood ratio transformation that were developed by Glanzer et al. (2009) to be used with unequal variance normal signal detection models. After the transformation has been applied, it results in log likelihood ratio distributions that are non-central chisquare in shape. We have had to modify the equations to consider cases in which the model has access to incomplete information about the study episode. For instance, consider a case in which items were either studied once (weak) or four times (strong). During the test phase, participants are tested on weak and strong targets in addition to lures. This leads to three memory strength distributions: one for lures, one for weak targets, and one for strong targets. If it's assumed that during weak target trials participants calculate the likelihood ratio using the weak target distribution as reference, the model has already presupposed memory of the test stimulus. Instead, under conditions in which mixed lists of items are studied, the expected strength in the likelihood ratio calculation is a distribution that reflects the average of the learning rate parameters corresponding to the weak and strong items[4]. Expected strengths were used in the likelihood calculations in the BCDMEM model (Dennis & Humphreys, 2001; Starns, White, & Ratcliff, 2010). The modified equation for the likelihood ratio transformation can be found in Appendix B. For all other parameter variations outside of the learning rate differences in mixed strength study lists, the expected parameter values are identical to the

[4]Another case where expected strengths would be employed is for modeling the effects of serial position. Items from different serial positions have different study-test lags, suggesting different values of the mean context match μ_{ss}. The expected context match at test could be constructed from the average of the context match parameters for each serial position.

actual parameter values of the memory strength distributions.

Predictions of the log likelihood ratio transformation can be seen in the bottom row of Figure 2. The left panel reveals that the transformation correctly produces the mirror ordering of the distributions (LN > HN > HO > LO). The model produces patterns that are quite similar to the variance account, but inspection of the 2AFC predictions in the right panel of Figure 2 reveals that the model is correctly able to predict null comparisons that are quite close to the experimental data (p(LO, HO) and p(HN, LN) > .5). The log likelihood ratio transformation is applied to all model predictions from this point on in the article.

Unequal Variance Between the Target and Lure Distributions of Memory Strength

The slope of the z-transformed ROC is almost uniformly less than one in the recognition memory literature (Egan, 1958; Glanzer, Kim, Hilford, & Adams, 1999; Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, McKoon, & Tindall, 1994). Within an SDT model, the common interpretation is the variability of the target distribution is greater than the lure distribution. When the distributions are normal in shape, the slope is the ratio of standard deviations $\sigma_{new}/\sigma_{old}$.

How can unequal variance be produced by our model? The variances of the target and lure distributions in item recognition are nearly identical except that for the case of targets, there are $l-1$ items in the item noise term and an additional self match term. If the self match variability exceeds the item noise for a single item (before being scaled by l), then targets will exhibit higher variability than lures. The model can accomplish this with sufficient values of either the item match variability (σ^2_{tt}) or context match variability (σ^2_{ss}). The effects of these variables can be seen in Figure 3 for both item recognition (top) and associative recognition (bottom): both the item match variability parameter σ^2_{tt} and context match variability parameter σ^2_{ss} were set to .02 and separately incremented to

higher values. As the values of these parameters are increased, one can see that the ratio of standard deviations decreases.

Why do item and context match variabilities produce unequal variance between the target and lure distributions? One might be inclined to think that the target item was seen in the list context, so variation in how the item was processed or variation in the match of the test context to the stored context should affect targets but not lures. However, both of these factors also affect the degree of match to previously stored memories. Inspection of Equations 13 and 14 reveals that the item mismatch in the item noise term is scaled by the context match and the context mismatch in the context noise term is scaled by the item match. Unequal variance is produced because in the self match term of Equation 13, there are more non-zero means contributing to the variance calculation (the learning rate r, the item match μ_{tt}, and the context match μ_{ss}). Thus, the simple answer is that the variability of the target distribution increases naturally with its mean, however the magnitude of this increase is modulated by the item and context match variability parameters. If the item and context match variability parameters exceed the item and context mismatch variability parameters, unequal variance that resembles the magnitude found in recognition memory experiments can be produced. As the item and context matching strengths decrease to zero, the ratio of standard deviations should approach 1. Ratcliff et al. (1994) found that with very low study times, the slope of the zROC is very close to 1. We constrain the values that these two parameters take by including ROC data, namely the dataset of DeCarlo (2007), in our model fit.

Much of the discussion about the source of unequal variance in recognition memory has focused on the hypothesis of Wixted (2007) that unequal variance arises due to variability in the strength of learning. That is, some items on the study list may be encoded with more strength than others, producing an additional source of variability for target items. The variability in learning strength hypothesis has been tested recently with mixed results (Koen & Yonelinas, 2010, 2013; Starns, Rotello, & Ratcliff, 2012; Jang,

Mickes, & Wixted, 2012). We would like to note that variability in the item match (which could arise from factors such as variability in the perceptual or semantic processing of the probe item) along with variability in the context match (which could arise from noise in either the contextual reinstatement process or the contextual drift process) are plausible contenders for sources of unequal variance that have not received attention in the literature. We do not mean this to imply that variability in the strength of learning is *not* responsible for unequal variance, but merely that unequal variance may reflect variability in several processes employed at both encoding and retrieval.

The Item Mismatch Variability Parameter and List Length/List Strength Predictions

As mentioned in the introduction, global matching models that only consider the role of item noise at retrieval predict detrimental effects of list length and list strength on recognition memory performance. From inspection of the item noise terms for item recognition (Equations 13 and 14) and associative recognition (Equations 16 and 17), one can see that item noise is produced by a multiplication of the item mismatch variability parameter σ_{ti}^2, the context mismatch variability parameter σ_{ss}^2, the learning rate r, and the number of items or pairs on the list l. The most critical of these parameters is the item mismatch variability. If this is set to zero, the entire item noise term is zero and no effect of list length or list strength is predicted. For positive values of the item mismatch variability parameter, increases in the number of list items l or the learning rate r increase the total item noise variance, producing poorer performance in conditions of higher list length or list strength, respectively.

These predictions can be seen in Figure 4 for both item recognition (top) and associative recognition (bottom). Depicted are two demonstrations for three different values of the item mismatch variability parameter σ_{ti}^2: 0, .02, and .04. The first demonstration is of a list length manipulation in which list length is varied between 1 and

80 items or pairs. Item noise increases linearly with increases in the list length for positive values of item mismatch variability. Consequently, performance decreases rapidly with increases in list length. When there is no item mismatch variability, item noise is zero and no effect of list length is predicted.

The list strength paradigm was simulated by using a 30 item list in item recognition and 30 pairs in associative recognition. Half of the items are baseline items that were studied with learning rate $r = 1.0$. The other half are interference items and were studied with learning rates varying between .05 and 2.5 and performance was assessed on the baseline items. As mentioned previously, for mixed lists of strong and weak items, the likelihood ratio computation compares items to a mixed distribution of strong and weak items. Thus, as the strength of the interference items is increased, the strength of the mixed distribution increases. This increase in the expected memorability of test items decreases the hit rates and false alarm rates of the non-strengthened items. This allows the model to predict the *strength based mirror effect* in item and associative recognition, which is where strengthening a set of items increases the hit rate of the strengthened items and simultaneously decreases the false alarm rate (Hirshman, 1995; Hockley & Niewiadomski, 2007; Stretch & Wixted, 1998b).

Due to the decrease in hit rates and false alarm rates that occur with increases in strength, the list strength effect can be more easily observed by observing how d' changes as the strength of the interference items and pairs is increased. When item mismatch variability is zero, there is no item noise and the strength of the interference items/pairs has no impact on performance. For positive values of item mismatch variability, performance degrades quickly as the strength of the interference items and pairs is increased[5].

[5]One should note that throughout this article, we follow convention in the recognition memory literature by visualizing performance in the model and the data using equal variance d' measures. When there is unequal variance between the signal and noise distributions, changes in bias result in changes in d'. Inspection of Figure 4 reveals that when the item mismatch variability parameter is zero, d' improves with list strength. This is due to the fact that both the log likelihood ratio distributions of targets and lures are shifted downward

One may also note that effects of both list length and list strength are smaller in associative recognition than in item recognition. That is because associative recognition involves the multiplication of two items instead of one, meaning that the item mismatch variability parameter σ_{ti}^2 is squared. When the values are less than one, σ_{ti}^2's influence on item noise will be larger for item recognition than for associative recognition. However, it is not the case that the item mismatch variability parameter has no visible effect on associative recognition predictions.

A number of researchers have conducted investigations using mixed lists of strong and weak pairs in associative recognition. On the test lists, rearranged pairs that came from both strong and weak pairs were presented and the false alarm rates were compared. A majority of investigations have found no difference between weak and strong rearranged pairs (e.g.: Buchler, Light, & Reder, 2008; Kelley & Wixted, 2001). As initially noted by Osth and Dennis (2014), the degree to which false alarm rates increase with strength in a mixed list is a consequence of item noise. To understand why, consider the partial match term in the equation for $\sigma_{rearr.}^2$:

$$2r_{assoc}^2(2\mu_{ss}^2\mu_{tt}^2\sigma_{ti}^2 + \mu_{ss}^2\mu_{tt}^2\sigma_{ti}^2 + \sigma_{ss}^2\mu_{tt}^2\sigma_{ti}^2 + \sigma_{ss}^2\sigma_{tt}^2\sigma_{ti}^2)$$

When item mismatch variability is zero, the interference from the two partial matches A-B and C-D to a rearranged pair A-D reduces to zero. When it is positive, the partial match term scales by the learning rate r_{assoc}, meaning that the two stored pairs exert greater interference on a rearranged pair cue as their strength is increased. This means that for a mixed list of strong and weak pairs, strong rearranged pairs suffer greater partial match interference when item mismatch variability is high.

The effect of item mismatch variability on mixed lists of strong and weak pairs can be seen in Figure 5. Three different levels of the parameter were compared ($\sigma_{ti}^2 = 0, .015, .03$) for a mixed list with 15 weak pairs studied with a learning rate $r_{assoc} = 1.0$ and 15 strong

---

as list strength is increased.

pairs were studied with $r_{assoc} = 2.5$. When item mismatch variability is zero, false alarm rates are equivalent between weak and strong rearranged pairs. However, as the item mismatch variability is increased, false alarm rates are considerably higher for strong rearranged pairs than for weak rearranged pairs. As we will discuss later, most investigations have found equivalent false alarm rates between weak and strong rearranged pairs (Cleary, Curran, & Greene, 2001; Kelley & Wixted, 2001; Osth & Dennis, 2014).

While null effects of list strength (e.g.: Ratcliff, Clark, & Shiffrin, 1990) and list length (e.g.: Dennis et al., 2008) are commonly found with word stimuli, novel non-linguistic stimuli such as fractal and face images have been found to be susceptible to effects of both list length (Kinnell & Dennis, 2012) and list strength (Osth et al., 2014; Norman, Tepe, Nyhus, & Curran, 2008). As we will demonstrate in the fits to our datasets, this can be accommodated by alotting separate item mismatch parameters to each stimulus class to reflect the idea that the item representations of each stimulus class may vary in their degree of inter-item similarity. Modeling the complete interference contributions between all the stimulus classes would require a matrix of item mismatch variability parameters that reflects item similarity within a given stimulus class and between the different stimulus classes. However, given that in all of the datasets included in our model fit only test one stimulus class, we simplify treatment by only considering within-class interference.

Background Noise

From our description of the item mismatch variability parameter, it might seem as if large values of that parameter will always produce positive effects of list length and list strength. However, this is not the case. As noted by Murdock and Kahana (1993a, 1993b), a large contribution from pre-experimental memories can be sufficient to drown out differences between two conditions that vary in their level of item noise, such as differences in list length or list strength.

The effect of the background noise parameter on list length and list strength predictions can be seen in Figure 6. The item mismatch variability parameter was set at .03, which was seen to produce relatively large item noise effects in Figure 4. Background noise (β) was varied for both item and associative recognition. The most obvious effect is that background noise degrades performance. However, as background noise increases, the increases in item noise with list length and list strength are relatively small compared to the interference already present in memory, and one can see that performance decreases at a smaller rate as list length or list strength are increased when the background noise present in memory is high.

While we have simplified the background noise contribution to a single parameter, the original parameterization describes it as follows:

$$n(\mu_{su}^2\sigma_{ti}^2 + \mu_{ti}^2\sigma_{su}^2 + \sigma_{su}^2\sigma_{ti}^2)$$

where n is the number of memories of the given stimulus class. Specifically, one can see that both n and the item mismatch variability parameter σ_{ti}^2 determine the total background noise. Given that the item mismatch variability parameter varies across stimulus classes and that each stimulus class likely varies in its number of entries in memory, it is also reasonable for the background noise to vary across stimulus classes. In our fits, β varies across stimulus classes but is constant across all other manipulations.

The demonstration that background noise can mask effects of list length and list strength is critical to our understanding of the sources of interference in recognition memory. Dennis and colleagues have previously argued that null effects of list length and list strength support the idea that there is no item noise in memory (Dennis & Humphreys, 2001; Osth & Dennis, 2014), whereas the presence of a significant contribution of background noise is a plausible alternative. Thus, null effects of list length and list strength do not necessitate a pure context noise model with no item noise. The model fits to a large number of experimental datasets allow us to distinguish between these possibilities.

The Model Fit

As previously mentioned, multiple possibilities in the parameter values can be responsible for recognition memory being impervious to manipulations of list length and list strength on recognition memory performance for words. For instance, null effects of list length and list strength could be predicted by a model with context noise as the sole source of interference (like in the BCDMEM model) or they could be predicted with a high ratio of background noise to item noise (like in the TODAM models). Similarly, overall levels of performance can reflect a strong degree of learning or low contributions of interference. The fact that multiple parameter combinations can qualitatively predict similar outcomes implies that the resulting parameter estimates of the model will not be independent but correlated with each other (e.g.: Turner, Sederberg, Brown, & Steyvers, 2013). Thus, the modeling exercise requires a robust model fitting procedure.

To properly measure the parameters required to estimate the interference contributions to recognition memory, we fit the model within a hierarchical Bayesian framework. The virtues of hierarchical Bayesian methods for fitting cognitive models (Lee, 2008, 2011; Lee & Vanpaemel, 2008; Pooley, Lee, & Shankle, 2011; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011) and in fitting data from recognition memory paradigms (Dennis et al., 2008; Pratte & Rouder, 2011; Pratte, Rouder, & Morey, 2010; Morey, Pratte, & Rouder, 2008; Turner, Dennis, & Van Zandt, 2013) have been well established. While traditional techniques such as minimizing the sum of squared deviations between the model's predictions and the data (approximate least squares) only provide point estimates for the model's parameters, one of the advantages of Bayesian analyses is that they quantify the uncertainty in the parameter estimates of the model as probability distributions over each parameter which are referred to as *posterior distributions*. Given that we are interested in quantifying the respective contributions of the interference contributions, properly measuring the uncertainty in these estimates is necessary.

Additionally, while it is common for modelers to aggregate across subjects and fit psychological models to only the group data, one problem is that fitting group estimates of the data often provides different parameter estimates than when the fits to the individual participants are averaged together (Estes & Maddox, 2005). Bayesian analyses can overcome this problem by usage of *hierarchical* models that jointly estimate the parameters of the group and the individual participants. This is accomplished by establishing *hyperparameters* that represent the group level parameters in combination with individual participant parameters. Each relevant model parameter (such as r, σ_{ti}^2, β, etc.) has its own set of hyperparameters, namely a mean and variance or precision parameters, that specify the prior distribution across all participants that each individual particiant's parameters are sampled from. A hierarchical fit provides posterior distributions on each model parameter for each participant along with posterior distributions for the hyperparameters of each model parameter that reflect the group-level parameters. Fitting individual participants in a modeling exercise such as this is critical, as there may be significant individual differences among participants in the magnitudes of the respective interference contributions.

As was previously mentioned, each interference source increases the variance of both the signal and noise distributions. In order to properly constrain the model parameters, it is required to include all of the relevant manipulations that affect the parameters of the model. It was for this reason that we included a large number of datasets, which include manipulations of list length, list strength, and also include mixed lists of strong and weak pairs in the associative recognition task, all of which constrain the estimates of item noise. We have additionally included two experiments that include manipulations of word frequency, which constrain the context noise parameter. Background noise is the remaining interference in the memory system and provides a constant source of noise that is unaffected by the experimental manipulations. In the next section, we will discuss the ten recognition memory datasets that are included in the model fit.

Datasets Included in the Model Fit

Rather than fit each dataset separately, we imposed significant constraint on the model by fitting all of the datasets simultaneously and constrained parameters across datasets. An advantage of the hierarchical Bayesian procedure is that individual participant parameters can be modeled while simultaneously constraining across different experiments or datasets by restricting the hyperparameters to be the same across datasets.

Hyperparameters were only allowed to vary across datasets where appropriate. For instance, the item mismatch variability parameter σ_{ti}^2 varies across stimulus classes, meaning that the Dennis et al. (2008) dataset, which employed words as stimuli, receives a different hyperdistribution for σ_{ti}^2 than for a dataset that used fractals as stimuli (such as Kinnell & Dennis, 2012, Experiment 2). However, other datasets that used words, such as the Osth and Dennis (2014) and the DeCarlo (2007) studies, share the same σ_{ti}^2 hyperdistribution that corresponds to word stimuli.

Here, we describe the datasets included in our model fit, their findings, the relevant hyperparameters they constrain, and briefly review the surrounding literature. A summary of all of the datasets used in the fitting can be seen in Table 2.

Table 2

Datasets included in the hierarchical Bayesian fit to the data.

Dataset	Task	Stim.	*N*	Resp.	Manip.
DC - Ex 1A	IR	Words	72	6 con.	Word frequency: LF and HF List length: 20 vs. 80 items
DLK	IR	Words	48	YN	Word frequency: LF and HF Unfilled vs. filled delay
KD - Ex 1	AR	Words	28	YN	List length: 24 vs. 96 pairs
KD - Ex 2	IR	Faces	39	YN	List length: 20 vs. 80 items
KD - Ex 3	IR	Fractals	32	YN	Same as above

Continued on next page

Table 2 - continued from previous page

Dataset	Task	Stim.	*N*	Resp.	Manip.
KD - Ex 4	IR	Scenes	40	YN	Same as above
ODK - Ex 1	IR	Fractals	88	YN	List strength: 32 items 1x vs. 16 items 1x 16 items 4x
ODK - Ex 2	IR	Faces	96	YN	Same as above
ODK - Ex 3	IR	Scenes	71	YN	Same as above
OD - Ex 1	AR	Words	80	YN	List strength: 32 pairs 1x vs. 16 pairs 1x 16 pairs 4x

Notes: Stim. = stimulus, *N* = number of participants, resp. = response type collected in the experiment, manip. = manipulations used in the experiment. DC = DeCarlo (2007), DLK = Dennis, Lee, & Kinnell (2008), KD = Kinnell & Dennis (2012), ODK = Osth, Dennis, & Kinnell (in press), OD = Osth & Dennis (2014), IR = item recognition, AR = associative recognition, YN = yes-no recognition, 6con = 6 point confidence rating scale.

The ROC in item recognition. As described earlier, the slope of the z-transformed ROC is almost uniformly less than one in the recognition memory literature (Egan, 1958; Glanzer et al., 1999; Heathcote, 2003; Ratcliff et al., 1992, 1994), a finding which has been used to advocate for the unequal variance signal detection model (Wixted, 2007). The ubiquity of the unequal variance interpretation of zROC slopes is further supported by evidence from response time models. Starns, Ratcliff, and McKoon (2012) found that the Ratcliff diffusion model (Ratcliff, 1978; Ratcliff, Van Zandt, & McKoon, 1999) could only fit the zROC slopes from a binary ROC paradigm (in which participants give yes or no responses, but the relative proportions of targets and lures are manipulated across conditions) if the variability of the drift rates for targets was larger than the variability for lures. More recently, Starns and Ratcliff (2014) fit a large number of recognition memory datasets that lacked complete ROC functions and found that a

diffusion model with higher drift rate variability for targets received better support than a diffusion model with equal variance across the drift rates for targets and lures.

As mentioned previously, unequal variance between the target and lure distributions can be produced in our model by choosing the appropriate values of the item and context match variability parameters σ_{tt}^2 and σ_{ss}^2. However, many of the datasets listed below lack the necessary data to constrain these parameters, in that they did not include manipulations of target vs. lure proportions or confidence ratings. It was for this reason that we additionally selected a relatively simple ROC experiment to constrain these parameters, namely Experiment 1A of DeCarlo (2007). This experiment tested native English participants for item recognition of both high and low frequency words using six point confidence ratings. To model this dataset, five hyperparameters were selected for the response criteria required to make the confidence ratings that were not employed in any of the other datasets. Differences between the word frequency classes were modeled by usage of the context mismatch variability parameter ρ: separate hyperparameters were used for both low and high frequency words. The presentation time for the stimuli was lower than for the other datasets (one second per stimulus), so this dataset was allotted its own hyperparameters for the learning rate r. Given that this experiment used immediate testing, the mean context match parameter μ_{ss} was fixed at 1.

The List Length Paradigm. Manipulations of the length of a study list have been a large constraint on models of memory, including models of recognition memory (Chappell & Humphreys, 1994; Clark & Gronlund, 1996; Dennis & Humphreys, 2001; Gillund & Shiffrin, 1984; Johns, Jones, & Mewhort, 2012; McClelland & Chappell, 1998; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997), free recall (G. D. A. Brown, Neath, & Chater, 2007; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Farrell, 2012; Polyn, Norman, & Kahana, 2009; Raaijmakers & Shiffrin, 1981; Sederberg, Howard, & Kahana, 2008), and serial recall (Botvinick & Plaut, 2006; G. D. A. Brown et al., 2000; Henson, 1998; Farrell, 2012; Lewandowsky & Murdock, 1989). In recognition memory, the earliest

demonstration of the detrimental effect of list length on recognition memory performance was found by Strong (1912). In the several decades that followed, the list length effect was replicated extensively in both item (Bowles & Glanzer, 1983; Cary & Reder, 2003; Gillund & Shiffrin, 1984; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991a; Nobel & Shiffrin, 2001; Ratcliff & Murdock, 1976; Underwood, 1978) and associative recognition (Clark & Hori, 1995; Nobel & Shiffrin, 2001) and has been used as support for pure item noise models of recognition memory (Clark & Gronlund, 1996; Gillund & Shiffrin, 1984).

However, as noted by Dennis and Humphreys (2001), a number of confounds exist in the previously published list length designs that may be artifactually causing a list length effect in recognition memory performance that we will briefly summarize here. First, if participants are tested immediately after completion of the study list, the average retention interval for items in the long list is longer than that of the short list. This confound can be overcome by equating the retention intervals across the two list length conditions by using a period of filler activity after the short study list is complete. Another confound is the fact that when immediate testing is used after a long list, participants may be more inclined to use a context representation that strongly favors the end-of-list items, rather than reinstating a list-wide context, a confound which can be overcome by having participants partake in additional filler task activity. An additional confound is that attention is likely to decrease through the duration of the study list, producing weaker encoding for the late list items relative to the early list items, a point which was first raised by Underwood (1978). This confound can be remedied by comparing items from equivalent serial positions from both the short and long list conditions.

When all of these confounds have been controlled, the investigations by Dennis and colleagues have found no effect of list length on performance when words are used as stimuli both in item recognition (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011) and associative recognition (Kinnell & Dennis, 2012, Experiment 1). Other investigations have found no effect of list length on performance in item recognition.

Schulman (1974) found no difference in 2AFC recognition performance between study lists of lengths 25, 50, and 100 items when the retention intervals and test positions were equated across the conditions. Jang and Huber (2008) found no difference between list lengths of 6 and 24 items in a 2AFC recognition task that interpolated between two lists that were later tested for free recall. Murnane and Shiffrin (1991a, Experiment 3) found no effect of list length on yes-no recognition performance when study-test lag was controlled.

While some might conclude that the controls employed in these list length paradigms might be sufficient to eliminate any effect of list length in recognition memory performance, this is not the case. Kinnell and Dennis (2012) tested novel non-linguistic stimuli, specifically images of fractals, faces, and natural scenes, in single item recognition using the same controls employed in the other investigations. Significant effects of list length were found for fractals and faces but no effect of list length was found for natural scenes. Kinnell and Dennis (2012) posited that the differences associated with the different stimulus classes may be attributed to different levels of item noise, with word stimuli and natural scenes being exempt from item noise at retrieval while faces and fractals suffer from item noise, possibly due to having more distributed item representations (we return to the issue of how different stimulus classes can suffer from different degrees of item noise in the General Discussion).

Included in the model fit are five experiments using the list length paradigm: the dataset of Dennis et al. (2008) along with the four experiments by Kinnell and Dennis (2012). All of the experiments use two list length conditions which all employ a 1:4 list length ratio from the short list to the long list. Additionally, all of the experiments only test the first 20 items (first 32 pairs for the associative recognition experiment) on the study list to make the test lists across the two list length conditions comparable. The five datasets comprise all of the stimuli that are employed in the model fit: words are employed in the study of Dennis et al. (2008) and the associative recognition experiment by Kinnell and Dennis (2012, Experiment 1). Faces, fractals, and natural scenes were compared using

separate experiments by Kinnell and Dennis (2012, Experiments 2-4, respectively).

The dataset of Dennis et al. (2008) was somewhat more extensive than the other datasets, as it also manipulated the length of a post study list delay period in addition to two different levels of word frequency. In the unfilled delay condition, recognition testing began immediately after the long list and 3 minutes after the end of the short list. In the filler condition, in contrast, recognition testing began an additional 8 minutes after the end of the long list and 11 minutes after the end of the short list. To capture the different levels of performance for each delay, separate hyperparameters for the mean context strength parameter μ_{ss} were allocated: one for the long list in the unfilled condition to allow for the possibility of poor contextual reinstatement, and another for both list length conditions in the filler task condition. To capture the effects of word frequency, the same context mismatch variability hyperparameters used in the fit to the dataset of DeCarlo (2007) were used to capture the effects of low and high frequency words in this dataset.

To test for the hypothesis that the different stimulus classes are subject to different degrees of item noise, separate hyperparameters for the item mismatch variability parameter σ^2_{ti} were allowed for each stimulus class. Osth et al. (2014) argued that the detrimental effects of list length and list strength observed with specific non-linguistic stimuli are too small to be accommodated by a pure item noise model and that these effects may be being mitigated by background noise from the memory system. For this reason, different hyperparameters for the background noise parameter β_{item} were allowed for each stimulus class. Additionally, separate background noise hyperparameters were allowed for word pairs in associative recognition to allow for the possibility that many more inter-item bindings are being stored in the co-occurrence tensor than item-context bindings in the occurrence matrix. Since the experiments of Kinnell and Dennis (2012) also used filler tasks that are of the same length as the filler task of Dennis et al. (2008), the same hyperparameters for the mean context strength were shared across all of these experiments.

The List Strength Paradigm. The null list strength effect was initially discovered by Ratcliff et al. (1990), who found that recognition performance for weak items was not harmed when they were accompanied by strong items on a study list and that strong items did not benefit from being accompanied by weak items relative to strong items. The null list strength effect was found in all seven of their experiments, regardless of whether the strengthening occurred via massed study, massed repetitions, or spaced repetitions. The list strength paradigm was extensively revisited in the two decades to follow, resulting in several replications of the null list strength effect in item recognition (Hirshman, 1995; Kahana, Rizzuto, & Schneider, 2005; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1994, 1992; Shiffrin, Huber, & Marinelli, 1995; Yonelinas, Hockley, & Murdock, 1992) and was recently demonstrated in associative recognition by Osth and Dennis (2014) using both yes/no and 2AFC testing. The finding of the null list strength effect has become a canonical constraint on recognition memory models (Chappell & Humphreys, 1994; Dennis & Humphreys, 2001; Johns et al., 2012; McClelland & Chappell, 1998; Norman & O'Reilly, 2003; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997).

The vast majority of the investigations that found no effect of list strength on recognition memory performance used single words or word pairs as study and test stimuli. Following the investigation of Kinnell and Dennis (2012), Osth et al. (2014) tested images of fractals, faces, and natural scenes for the presence of a list strength effect while simultaneously employing the list length controls advocated by Dennis and colleagues[6]. A significant list strength effect was found for fractals with both yes/no and 2AFC testing while null effects of list strength were found for faces and scenes (although a significant list strength effect was found using artificial faces by Norman et al., 2008). Osth et al. (2014) used this finding to support the hypothesis of Kinnell and Dennis (2012) that novel

[6]As noted by Osth et al. (2014), while the controls for differences in retention interval, attention, and contextual reinstatement are not commonly applied in list strength paradigms, they can similarly contribute to the artifactual finding of a list strength effect.

non-linguistic stimuli may be more susceptible to the effects of item noise.

An additional regularity of the list strength paradigm is the strength based mirror effect, in which strengthening a set of list items results in higher hit rates for those items along with a lower false alarm rate. Hirshman (1995) found this pattern to be ubiquitous in the early published list strength paradigms, and several investigations have replicated the effect in both item recognition (Criss, 2006, 2009, 2010; Hockley & Niewiadomski, 2007; Singer, 2009; Starns, Ratcliff, & White, 2012; Starns et al., 2010; Starns, White, & Ratcliff, 2012; Stretch & Wixted, 1998b) and associative recognition (Clark & Shiffrin, 1992; Hockley & Niewiadomski, 2007; Osth & Dennis, 2014). Furthermore, all of the non-linguistic stimuli in the study by Osth et al. (2014) exhibited a strength based mirror effect.

Included in the model fit are four experiments using the list strength paradigm: the datasets of Osth et al. (2014) and Osth and Dennis (2014). All of these experiments compared lists of 32 unique items or pairs of items across two conditions: a pure weak condition where all items or pairs were presented once, along with a mixed list condition where half of the items or pairs were presented once and the other half were presented four times. To make the tests of both list conditions comparable, all repetitions in these studies occurred after all of the unique items were presented once. Fractals, faces, and natural scenes were the stimuli of Osth et al. (2014, Experiments 1-3, respectively) and word pairs were employed in the study of Osth and Dennis (2014). Only the yes-no data from these investigations were employed.

In all experiments, hit rates for strong items or pairs greatly exceeded those of weak items or pairs. This was accommodated by allocating a separate learning rate r_{item} or r_{assoc} for the items or pairs that were presented four times. To guarantee that learning rates for strong items exceeded those for weak items, the samples of the strong learning rates were added to the weak learning rates. The value of the strong learning rates along with the value of the item mismatch variability parameter influence the magnitude of the list

strength effect.

One aspect of the associative recognition dataset that distinguishes it from the item recognition datasets is that the test lists in the mixed lists comprised rearranged pairs constructed from both weak (once presented) and strong (four times presented) pairs. Several studies have made such a comparison and found equivalent false alarm rates to both weak and strong pairs in young adults (Buchler et al., 2008; Cleary et al., 2001; Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Kelley & Wixted, 2001; Mickes, Johnson, & Wixted, 2010; Osth & Dennis, 2014)[7], which can be considered a broken within-list strength based mirror effect, although some studies have found reduced false alarm rates for strong pairs under specific conditions such as delayed responding or substantial strength differences between weak and strong pairs (Light et al., 2004; Malmberg & Xu, 2007; Xu & Malmberg, 2007). A unique aspect of the Osth and Dennis dataset is that both an intact and broken strength based mirror effect were observed: false alarm rates were lower in mixed lists than in pure weak lists (an intact across-list strength based mirror effect), but false alarm rates to weak and strong pairs were equivalent in the mixed list (a broken within-list strength based mirror effect). As was previously mentioned, the false alarm rates between weak and strong pairs is an additional constraint on the item mismatch variability parameter σ_{ti}^2.

A Note on Response Criteria

In the Turner, Dennis, and Van Zandt (2013) model fit, the criteria of the model were fixed at 0.0, which reflects the point on the log likelihood ratio scale where an item is equally likely to be a target or a lure. Nonetheless, some recent evidence has suggested

[7]Higher false alarm rates to strong pairs have been observed in older adults (Buchler, Faunce, Light, Gottfredson, & Reder, 2011; Light, Patterson, Chung, & Healy, 2004). While one possibility for this result is that older adults suffer from more item noise at retrieval, another possibility is that they rely more on item information in the associative recognition task than in younger adults, possibly due to a poorer ability to encode and use associations (Naveh-Benjamin, 2000).

that there is substantial variability in criteria across participants. Aminoff et al. (2012) found reliable individual differences in criteria across participants and found they were predicted by such psychometric variables as personality traits and affect. Kantner and Lindsay (2012) found that individual differences in response criteria were stable across recognition tests that were up to a week apart, although they varied somewhat across different stimulus materials.

Thus, we allowed for individual differences in the decision criterion parameter by having each participant's criterion sampled from a set of hyperparameters. Additionally, given the differences among different experiments in terms of the perceived difficulty of the stimulus materials, we have allowed different hyperparameters for each experiment to be used. Criteria were not allowed to vary across conditions or across trials within a given experiment.

A Note on Participant Exclusion

The studies of Kinnell and Dennis (2012), Osth et al. (2014), and Osth and Dennis (2014) all used experimental parameters that were quite similar to each other. Nonetheless, the studies of Osth and Dennis and Osth et al. excluded participants who exhibited d' that was zero or less in one of their experimental conditions, as these were likely participants who were not properly following instructions. As a consequence, the performance of the groups in these studies was noticeably higher than those of the Kinnell and Dennis experiments. To ensure that the data from the studies are comparable to each other, we used the same exclusion criteria on the experiments from Kinnell and Dennis. This resulted in the exclusion of twelve participants from Experiment 1, one participant from Experiment 2, and eight participants from Experiment 4. No participants from Experiment 3 met these exclusion criteria.

The Hierarchical Bayesian Model

It is quite common to express hierarchical Bayesian analyses as graphical models (Jordan, 2004). However, due to the large number of datasets being used in the fit in addition to the fact that several parameters are constrained across the fits, a complete graphical model would be too large and cumbersome to be useful to the reader. Instead, we have presented a general graphical model in Figure 7 that describes the parameters of the model whereas descriptions of the entire set of hyperparameters and the datasets and conditions to which they apply can be found in Table 3.

One of the advantages of the Bayesian approach is the ability to restrict the parameter space via a specified prior distribution to reflect a priori beliefs about how the parameter is distributed (Vanpaemel & Lee, 2012). However, given that this model has not been previously fit to data, we use nearly non-informative priors instead which place approximately equal likelihood over all values in the parameter space. For participant parameters that are bounded between 0 and 1 (r and μ_{ss}), parameters were sampled from beta distributions that were reparameterized in terms of the mean (λ) and variance (ν) parameters of the beta distribution. This reparameterization is achieved from the initial parameters α and β by setting $\alpha = \lambda\nu$ and $\beta = (1 - \nu)\lambda$. Prior distributions on these parameters were as follows:

$$\lambda \sim Beta(.5, 2)$$

$$\nu \sim InverseGamma(.1, .1)$$

For parameters that are bounded between 0 and ∞ ($\sigma^2_{tt}, \sigma^2_{ss}, \sigma^2_{ti}, \rho$, and β), participant parameters were sampled from lognormal distributions. For parameters that are bounded between $-\infty$ and ∞ (the criterion parameters), participant parameters were sampled from normal distributions. Prior distributions on the mean (ω) and precision (ξ) of the normal

and lognormal distributions were specified as follows:

$$\omega \sim Normal(0, .001)$$

$$\xi \sim InverseGamma(.1, .1)$$

All of the model's parameters were used to calculate the means and variances of the memory strength distributions according to Equations 12, 13, and 14 for item recognition and Equations 15, 16, and 17 for associative recognition. The memory strength distributions were then converted to log likelihood ratio distributions using the equations in Appendix B. For the experiments that used the yes-no response procedure, hit (h) and false alarm (f) rates were calculated by taking the area above the response criterion. Hit (H) and false alarm (F) count predictions for each participant i in a given condition j in experiment k were sampled from a binomial distribution:

$$H_{i,j,k} \sim Binomial(h_{i,j,k}, T_{j,k})$$

$$F_{i,j,k} \sim Binomial(h_{i,j,k}, L_{j,k})$$

where T and L refer to the number of target and lure trials. For the DeCarlo (2007) experiment which utilized confidence ratings, h and f were calculated for each confidence category c by calculating the area between the criteria for the middle responses, the area above the highest criterion for the highest confidence rating, and the area below the lowest criterion for the lowest confidence rating. H and F predictions for each confidence category were sampled from a multinomial distribution:

$$H_{c1,i}, ..., H_{c6,i} \sim Multinomial(h_{c1,i}, ..., f_{c6,i}, T)$$

$$F_{c1,i}, ..., H_{c6,i} \sim Multinomial(f_{c1,i}, ..., f_{c6,i}, L)$$

where $c1$ is the lowest confidence rating and $c6$ is the highest confidence rating.

The hierarchical model was fit using JAGS software (Plummer, 2003). The data were fit to all 594 participants from each of the ten aforementioned datasets simultaneously. The

data from each participant were the raw response counts for targets and lures in each condition. The results of the model fit are based on 32 chains, each consisting of 10,000 samples after 4,000 burn-in samples were discarded. Chains were visually checked for convergence.

Table 3

All hyperparameters included in the hierarchical model.

Param.	Number	Cond.	Datasets
r_{item}	1	Words (1 sec.)	DC
	2	Words (3 sec.)	DLK
	3	Fractals 1x	KD Ex 2; ODK Ex 1
	4	Fractals 4x	ODK Ex 1 (Mixed list cond.)
	5	Faces 1x	KD Ex 3; ODK Ex 2
	6	Faces 4x	ODK Ex 2 (Mixed list cond.)
	7	Scenes 1x	KD Exp 4; ODK Ex 3
	8	Scenes 4x	ODK Ex 3 (Mixed list cond.)
r_{assoc}	1	Pairs 1x	KD Ex 1; OD
	2	Pairs 4x	OD (Mixed list cond.)
μ_{ss}	1	Short delays (3.5 min)	ODK Ex 1, 2, 3; OD
	2	Long delays (8 min)	DLK; KD Ex 1, 2, 3, 4
	3	Long list, no filler	DLK
σ^2_{tt}	1	All	All datasets
σ^2_{ss}	2	All	All datasets
σ^2_{ti}	1	Words	DLK; DC; KD Ex 1; OD
	2	Fractals	KD Ex 2; ODK Ex 1
	3	Faces	KD Ex 3; ODK Ex 2
	4	Scenes	KD Ex 4; ODK Ex 3

Continued on next page

Table 3 - continued from previous page

Param.	Number	Cond.	Datasets
ρ	1	LF words	DLK; DC
	2	HF words	DLK; DC
β_{item}	1	Words	DLK; DC
	2	Fractals	KD Ex 2; ODK Ex 1
	3	Faces	KD Ex 3; ODK Ex 2
	4	Scenes	KD Ex 4; ODK Ex 3
β_{assoc}	1	Pairs	KD Ex 1, OD
	1-5	Confidence	DC
Φ	6		DLK
	7		KD Ex 1
	8		KD Ex 2
	9		KD Ex 3
	10		KD Ex 4
	11		ODK Ex 1
	12		ODK Ex 2
	13		ODK Ex 3
	14		OD Ex 1

Notes: Param. = parameter, cond. = condition, DC = DeCarlo (2007), DLK = Dennis, Lee, & Kinnell (2008), KD = Kinnell & Dennis (2012), ODK = Osth, Dennis, & Kinnell (in press), OD = Osth & Dennis (2014), 1x = once presented, 4x = four times presented. Note that each parameter receives its own mean and variance/precision parameter - see the text for details.

Analysis of the Model Fit

In our presentatation of the fit of the model to the data, we present both group level predictions and predictions for the individual participants. To assess the goodness of fit at the group level, we follow convention established in the recognition memory literature and restrict analyses to hit and false alarm rates and d'. Group level predictions were derived from the means of the hyperparameters which correspond to group-level estimates of the relevant parameters. For parameters that were lognormally distributed, the hypermean ω parameters were transformed as e^{ω}, which is both the geometric mean and median of the lognormal distribution[8].

Space precludes depiction of how the set of 594 individual participant parameters of the model were able to fit the data. Instead, we depict the individual hit and false alarm counts from each participant along with the model's posterior predictive distribution. Unlike the group level predictions, the posterior predictive distribution uses the entire hyperdistribution: hit and false alarm predictions are generated for each sample of the mean and variance/precision parameters. For analysis of the individual participants' interference contributions, individual participant parameters were used. The predictions and data are depicted on a scatterplot with hits on the y-axis and false alarms on the x-axis.

Where necessary, inferential statistics were performed on model parameters by taking the difference between the means of the hyperparameters and evaluating the proportion of samples that are above zero, which measures the probability of a difference between the two model parameters. Additionally, all density estimates on the posterior distributions were performed using Gaussian kernel density estimation.

[8]The arithmetic mean of the lognormal distribution is $e^{\omega+1/2\xi^2}$. We preferred usage of the geometric mean/median e^{ω} due to the strong degree of skew in the lognormal distribution, which makes the arithmetic mean a worse measure of central tendency.

Parameter Estimates

Posterior distributions of the group means for each parameter can be seen in Figure 8. Several parameters largely conform to expectations. The learning rate r is highest in conditions of strong performance: repeated items in the list strength experiments have the highest learning rates, and the learning rate is higher in the Dennis et al. (2008) dataset than the DeCarlo (2007) dataset, which is sensible given the higher presentation time in the former experiment. Learning rates were also higher for better performing stimulus classes, with words, word pairs, and scenes showing the highest learning rates and fractals and faces exhibiting the poorest learning. The context match parameter μ_{ss} was expected to vary with study-test delay: estimates of this parameter varied by the duration of filler activity as predicted, with the lowest values seen for the 8 minute filler activity and higher values for the 3.5 minute filler activity and the long list, no filler condition of Dennis et al. (2008) to reflect poor contextual reinstatement. The context mismatch parameter ρ, which influences the magnitude of context noise, varied as expected with higher values for high frequency than low frequency words to reflect their greater occurrences in memory. A slightly negative bias in the criteria (Φ) can be seen for all stimulus classes except for scenes, which exhibit a more conservative bias.

As was previously mentioned, Osth et al. (2014) attributed the small item noise effects seen for nonlinguistic stimuli to a higher degree of both item and background noise than word stimuli would exhibit, and the resulting parameter estimates conformed to these predictions. Both the item mismatch parameter σ_{ti}^2 and the background noise parameter β vary by stimulus class largely as predicted. Difference distributions for these parameters can be seen in Figure 9. The highest values for the item mismatch variability parameter are for fractals and faces. For fractals, 95.4% of the fractals minus words difference distribution lies above zero. For the faces minus words comparison, 99.9% of the difference distribution lies above zero. Images of natural scenes do not appear to differ significantly from words, as only 33.7% of the scenes minus words difference distribution lies above zero.

For the background noise parameter, fractals, scenes, and pairs differ significantly from words, with 99.9%, 99.2%, and 100% of the area of the respective difference distribution lying above zero. While faces appear to exhibit more background noise than words, 82.6% of the area of the faces minus words difference distribution lies above zero. Further comparison of the complete interference estimates for each stimulus class can be seen in the section Interference Contributions.

The Model Fit

While the primary purpose of fitting the model to data was to measure the magnitudes of the different interference contributions, interpretation of the model parameters is also reliant on the model achieving a good quantitative fit to the data. In this section, we present the model's predictions alongside both the group and individual participant data. For the group data, rather than compare the hit and false alarm rates generated from frequentist methods, we separately estimated the hit and false alarm rates for each dataset using simple hierarchical models and depict the predicted rates from the group mean parameters. Details of how the binomial and multinomial rates were estimated can be seen in Appendix C.

Word Frequency and Confidence Ratings: Fit to DeCarlo (2007). The fit to the DeCarlo (2007) dataset can be seen in Figure 10 for the standard ROC and Figure 11 for the z-transformed ROC. To generate a density estimate of the ROC function, the density was estimated on all of the points of the confidence based ROC simultaneously. Inspection of the graphs reveals that the model exhibits a close correspondence to the experimental data, with better predicted performance for LF words than HF words.

The model misses slightly on the zROC slopes: predicted zROC slopes for the median zROC points for the data are .81 and .83 for LF and HF words whereas the model produced zROC slopes of .91 and .95 for LF and HF words. This may be because inspection of Figure 11 reveals that the data's zROCs for HF words are slightly curvilinear,

whereas the model is only capable of producing linear zROC slopes. Curvilinear zROC slopes have been attributed to various factors, such as the presence of recollection (Yonelinas, 1994), probability mixtures of encoded and non-encoded items (DeCarlo, 2002), and decision noise (Ratcliff & Starns, 2009). We would like to emphasize that the primary purpose of including this dataset in our omnibus fit was not to discriminate between different theoretical explanations of ROC functions, but to constrain the parameters of the model that are principally responsible for producing unequal variance between the target and lure distributions, namely the variance in the item and context match parameters σ_{tt}^2 and σ_{ss}^2. Inspection of Figure 8 reveals that the parameters are well constrained by the data, as their posterior distributions appear within only a limited range of their prior distributions, which are broad and non-informative.

Individual participant data along with the model's posterior predictive distribution can be seen in Figure 12. One can see that the density of the model's predictions closely follows the density of the individual participant responses.

List Length, Word Frequency, and Study-Test Delay: Fit to Dennis, Lee, and Kinnell (2008). The fit to the group data of Dennis et al. (2008) can be seen in Figure 13. For this dataset and all remaining datasets, the density estimates of the posterior distributions of both the rates estimated from the data along with the model's predicted rates are depicted using teardrop plots, which are vertical depictions of the posterior distribution. A teardrop plot is constructed by plotting a posterior distribution sideways for both the left and the right. Areas of the teardrop plots with greater width indicate higher density regions of the posterior distribution.

The model achieves an excellent fit to the data, with the posterior distributions of the model's predictions aligning very closely with the posterior of the group data. As mentioned previously, this dataset exhibited a small list length effect in the no filler condition while there was no effect of list length in the filler condition, similar to what is seen in the data. Dennis et al. (2008) argued that when study lists are immediately

followed by a test list, participants might be more likely to use an end-of-list context than reinstate a list-wide context. We allowed for this possibility by allowing a separate set of μ_{ss} hyperparameters for the long list in the no filler condition. While inspection of Figure 8 revealed that the poor contextual reinstatement μ_{ss} parameter ended up quite high, it was still sufficient to allow the model to predict a small effect of list length in the no filler condition.

Due to the low estimate of the item mismatch variance parameter σ^2_{ti} for word stimuli, the model predicts virtually no effect of list length in the filler condition. The model also predicts the word frequency mirror effect in the data, predicting both higher hit rates and lower false alarm rates for low frequency words. The performance decrement of delayed testing in the 8 minute filler condition is also addressed by the model, although the magnitude of the performance decrement on both the data and the model's predictions appears to be small.

One should also note that several of the parameters used in the fit to this dataset were constrained across other datasets. The context mismatch variability parameters ρ for low and high frequency words, along with the background noise parameter β and item mismatch variability parameter for word stimuli σ^2_{ti} were also employed in the fit to the data from DeCarlo (2007). The item mismatch variability parameter for word stimuli was also used in the fits to the experiments that employ word pairs, namely the associative recognition experiments that manipulate list length and list strength conducted by Kinnell and Dennis (2012) and Osth and Dennis (2014).

Individual participant data along with the model's posterior predictive distribution can be seen in Figure 14. One can see that the density of the model's predictions closely follows the density of the data.

List Length and List Strength with Non-Linguistic Stimuli: Fit to Kinnell and Dennis (2012) and Osth et al. (2014)

Fits to all of the datasets with non-linguistic stimuli, specifically fractals, scenes, and faces, can be seen in Figure 15. This includes the list length experiments conducted by Kinnell and Dennis (2012) and the list strength experiments conducted by Osth et al. (2014). The fit to the data is quite good, with the model's predictions falling within the posterior distribution of the data for every comparison. Inspection of the list strength data reveals that the model is adept at predicting the large strength based mirror effect for fractals and for faces, with much lower false alarm rates being predicted for the mixed condition than for the pure weak condition.

The magnitude of the list length and list strength effects is difficult to evaluate on the basis of hit rates and false alarm rates alone. d' estimates can be seen in Figure 16. For fractals and faces, both the data and the model reveal lower d' in the long list and mixed list conditions relative to the short list and pure weak conditions. As was previously mentioned, the detrimental effects of list length and list strength are quite small in both the data and the model's predictions. For scenes, there is no effect of list length or list strength. These predicted effects correspond to the differences in item mismatch variability depicted in Figure 8, as list length and list strength effects are predicted for the stimuli with the highest values of item mismatch variability.

Individual participant data along with the model's posterior predictive distribution for both the list length and list strength paradigms can be seen in Figures 17 and 18. The fit appears to be quite good.

List Length and List Strength with Word Pairs in Associative Recognition: Fit to Kinnell and Dennis (2012) and Osth and Dennis (2014)

Fits to the datasets that employ the associative recognition task with word pairs can be seen in Figure 19. This includes the list length experiment of Kinnell and Dennis (2012,

Experiment 1) and the experiment of Osth and Dennnis (2014) that utilized yes/no responding. The fit to the data is good, with the model predictions falling within the posterior distributions of the data. Nonetheless, the fit to the list length experiment is somewhat worse than the fit to the list strength experiment. This is in part because the performance of the participants appear to be much poorer in the list length experiment than in the list strength experiment despite the similar experimental parameters employed in both experiments. d' estimates can be seen in Figure 20. The model predicts no effect of either list length or list strength on associative recognition performance. While performance appears to be somewhat poorer in the long list, this effect was found to not be significant by Kinnell and Dennis (2012) in their analyses.

As mentioned previously, the item mismatch variability parameter σ_{ti}^2 is not just constrained by the manipulations of list length and list strength, but higher values of σ_{ti}^2 predict higher false alarm rates to strong rearraned pairs than weak rearranged pairs in mixed lists. Inspection of Figure 19 reveals that the model predicts nearly equivalent false alarm rates between weak and strong rearranged pairs, like in the data. Thus, the model is able to simultaneously predict the cross-list strength based mirror effect (lower FAR in the mixed list than in the pure weak list) as well as the broken within-list strength based mirror effect (weak FAR = strong FAR). This is because the strength estimates used in the likelihood ratio calculation are list-wide: when the strength of a list changes, the strength estimates change. However, on a given test list, the strength estimates are not permitted to change across trials. This is conceptually quite similar to the hypotheses of Stretch and Wixted (1998b) and Hockley and Niewiadomski (2007), who posited that criterion shifts occur across lists but stay relatively constant within a test list.

Individual participant data along with the model's posterior predictive distribution for both the list length and list strength paradigms can be seen in Figures 17 and 18.

Interference Contributions

To evaluate the contributions from each interference component, the means of the hyperparameters were used to calculate each interference component according to Equations 14 and 17. Because the variance of the self match is less relevant than the other contributions, only the components of the lure distribution were used in the calculation. For word pairs in associative recognition, we combined the partial match and item noise terms from Equation 17 due to the shared influence of the item mismatch variability parameter on their predicted magnitudes. Density estimates of the item noise, context noise, and background noise calculated for each dataset in the fit can be seen in Figure 22. Given that context noise was only measured for single word stimuli, context noise is only present in the fits to DeCarlo (2007) and Dennis et al. (2008).

Figure 22 depicts the total interference contributions in order from smallest to largest. For the item noise estimates, only the conditions with the highest item noise were used (long lists in list length manipulations and mixed lists in list strength manipulations). One can see that the item noise contributions for words and scenes along with context noise for low frequency words rank as the smallest interference contributions out of all of the interference contributions. Figure 22 also confirms that the differences in the item mismatch variability parameters across stimulus classes extends to the total item noise as well. Item noise is significantly higher for fractals and faces than for single words, word pairs, and scenes. Additionally, the largest interference estimates appear to be background noise and context noise for high frequency words. Despite the fact that background noise estimates were low for single words, they appear to be quite large relative to the other interference contributions. Item noise, in contrast, does not appear to dominate the other interference contributions in any of the datasets.

To elucidate the relative contributions of each interference component, proportions of total interference were calculated for each interference term. For each dataset, the condition with the highest item noise was used (long lists in the list length experiments

and mixed lists in the list strength experiments). Given that the dataset of (DeCarlo, 2007) used shorter lists (70 items) than the longest lists in the (Dennis et al., 2008) dataset (80 items), only the dataset of the Dennis et al. fit was used in this analysis. Additionally, given that only one context noise term contributes at retrieval, separate analyses were performed for the low and high frequency words.

Proportions of total interference can be seen in Figure 23, which contains bar plots of the median proportion of total interference for each interference component along with the 95% highest density interval (HDI). For single words, word pairs, and scenes, item noise is extremely close to zero in its contribution to the total interference. For fractals and scenes, item noise occupies a much greater proportion of the total interference. Nonetheless, for fractals and faces, background noise occupied a significantly greater proportion of interference than item noise in the list length datasets, with 99.5% and 97.8% of the background noise minus item noise difference distribution above zero for fractals and faces, respectively. The differences are more ambiguous for the list strength datasets, with 94.2% and 86.1% of the background noise minus item noise difference distribution above zero for fractals and faces. For scenes and word pairs, background noise occupies virtually all of the total interference.

Converging evidence was found in an analysis of the interference contributions of the individual participants. For each participant, difference distributions were constructed between the item noise of the highest item noise condition and the other interference contributions. Following statistical conventions, differences among the interference contributions were deemed significant if 95% of the area of the difference distribution lied above zero. For words in the Dennis et al. dataset, all of the participants exhibited significantly higher background noise than item noise and significantly higher context noise for high frequency words than item noise. For both word pairs and scenes, all of the participants exhibited higher background noise than item noise in both the list length and list strength datasets. For fractals, one participant exhibited significantly higher

background noise than item noise for the list length dataset. Otherwise, for both fractals and faces, none of the participants exhibited a dominant interference contribution.

One of the surprising findings of this analysis is that context noise is virtually zero for low frequency words (median of .03% of total interference), with background noise occupying nearly all of the interference (median: 98.7%). For high frequency words, roughly equivalent levels of context noise and background noise are present. How could low frequency words exhibit such little context noise? In our fits, we collapsed across the number of occurrences of a word (m) and the variability in the similarity to previous contexts (σ^2_{su}). One possibility is that contexts are quite dissimilar to previous contexts, meaning that the true value of σ^2_{su} is quite low. High frequency words may suffer from considerable interference not because of the overlap among contexts, but due to their frequent exposures (a high value of m).

While we have reported that the model's fit is quite good, a model can often times fit well because it is overfitting the data (Pitt & Myung, 2002). One way to ensure that a model is capturing the underlying structure of the data is to fit the model to only a sample of the total data and evaluate how well it performs on the remaining data, a technique called cross validation. We performed a k fold cross validation procedure, which has been shown to outperform the leave-one-out cross validation (LOOCV) method (Arlot & Celisse, 2010). In the k folds procedure, the data is equally divided into k sections, or folds, and the model is independently fit to each fold. For each fold, the model's generalizability was evaluated by comparing the model's predictions to the withheld data. Not only was the model well able to fit the withheld data, but parameter estimates were consistent with those derived from the main fit. The description and results of the cross validation procedure can be seen in Appendix D.

General Discussion

When cues are globally matched against the contents of memory the output of the retrieval process can be characterized on the basis of matches and mismatches to the item and context cues employed at retrieval. We fit a global memory model based on the tensor model of Humphreys, Bain, and Pike (1989) that directly parameterizes the matches and mismatches to the item and context cues to ten recognition memory datasets. The model allows for an analytic estimation of the contributions of item noise, context noise, and background noise that directly follow from the parameters of the model. While a model is made more flexible by inclusion of all possible interference sources, the fact that the different stimulus classes differed along these dimensions in psychologically meaningful ways supports taking such a comprehensive approach. Moreoever, the parameters of our model were constrained by the manipulations of strength, list length, list strength, word frequency, study-test delay, along with the different stimulus classes. Maximum constraint was imposed on the model by constraining several of the model's parameters to be constant across several of the datasets in the fit, which contained a total of 594 participants.

Resulting parameter estimates derived from a hierarchical Bayesian analysis revealed that item noise plays a rather small role in retrieval, although the magnitude of its influence depends on the stimulus class. Estimates of item noise were jointly constrained by the manipulations of list length, list strength, as well as the simultaneous testing of weak and strong pairs in a mixed list. For words, word pairs, and scenes, item noise is extremely close to zero and interference stems entirely from pre-experimental sources, namely context noise and background noise. For fractals and faces, item noise is much larger, although background noise appears to be the dominant source of interference for these stimulus classes in some comparisons (the list length datasets). None of the analyses revealed a dominant influence of item noise in any of the stimulus classes or individual participants. These results were particularly surprising because the contributions of background noise have generally been ignored by models of episodic recognition. In the

following sections, we discuss the implications for these parameter estimates in the recognition memory literature, such as what the parameter estimates imply about the underlying representations of linguistic and non-linguistic stimuli and comparisons to previous investigations which have argued for higher magnitudes of item noise.

Plausibility of Low Item Noise

While the results of our model fit demonstrate that item noise makes a relatively small contribution to the total interference in recognition memory tasks, it does not specify what the vector representations of the items are that would result in low item noise. It has been argued that no interference among the items can be exhibited when item representations are orthogonal to each other (Dennis & Humphreys, 2001; Osth & Dennis, 2014). At a psychological level, this would imply that item representations are dissimilar and share no features with each other. The results of our fitting imply that there is a non-zero contribution of item noise at retrieval, which rejects the notion of orthogonal item representations. Nonetheless, a close approximation would be employing relatively sparse item representations, meaning that they are not completely orthogonal but exhibit minimal overlap with each other.

Some might find such an idea implausible, especially given that words in recognition memory tasks are often perceived as similar to each other in meaning, phonology, and surface form. How then could their representations be dissimilar? A number of theories of hippocampal function describe the function of the hippocampus as creating sparse high dimensional representations from overlapping inputs (Kumaran & McClelland, 2012; Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995; Treves & Rolls, 1992; Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001), allowing the hippocampus to exhibit fast learning, discriminate among highly similar novel inputs, and minimize interference, all qualities that are critical for performance on episodic memory tasks. Tasks which require the similarity among the items to be emphasized can employ

the distributed representations in the neocortex that have been developed from experience, which is a central component of the complementary learning systems (CLS) theory (McClelland et al., 1995).

Another possible interpretation is that item representations are multidimensional and that their dimensions can be weighted based on their relevance to a cognitive task, as proposed by the generalized context model (GCM: Nosofsky, 1986, 1991). In this regard, dimensions among the items in a recognition memory task that emphasize their similarity, such as their shared semantics or surface appearance, may be de-weighted for optimal discrimination of old from new items. This approach yields similar interpretations as to the hippocampal theories, in that representations of items are not fixed, but adapt to the task faced by the participant. Thus, the results of our model fitting do not imply that word representations are dissimilar in all cognitive tasks, merely that they are dissimilar in episodic memory.

Another possible objection to the finding of minimal item noise in recognition memory is that in the free recall task, manipulations of list length (Murdock, 1962; Roberts, 1972; Ward, 2002) and list strength (Malmberg & Shiffrin, 2005; Ratcliff et al., 1990; Tulving & Hastie, 1972) exhibit robust decrements on recall performance. How could a memory system which exhibits minimal interference among the items in recognition memory exhibit such strong competition among the items in a free recall task? As it turns out, the majority of current free recall models exhibit competition among the items not because of overlap in their item representations, but due to usage of a sampling with replacement memory search process (Davelaar, 2007; Wixted & Rohrer, 1994).

In resampling models, the context cue initiates the probabilistic sampling of items that are most strongly bound to the list context. As recall progresses, previously recalled items are not removed from the set of recall candidates. Instead, they can continue to be sampled by the search process instead of items that have not yet been recalled. List length effects fall naturally from resampling models as longer lists contain more candidates to be

output, decreasing the probability that any particular item will be sampled. It has been demonstrated by both Wixted and Rohrer (1994) and Davelaar (2007) that resampling allows for the prediction of both list length effects and the increase of inter-response times as recall proceeds (Murdock & Okada, 1970). Additionally, other current resampling models of free recall with specified item and context representations, such as the model of Davelaar et al. (2005) and the temporal context model (Howard & Kahana, 2002; Sederberg et al., 2008), employ orthogonal item representations and predict detrimental effects of list length on free recall performance, which demonstrates that similarity among the list items is not necessary to predict competition among the items in free recall.

While TCM and the model of Davelaar et al. have not simulated the list strength paradigm, resampling provides an intuitive explanation: increasing the strength of a subset of list items increases their sampling probability and decreases the sampling probability of the non-strengthened items (Wixted, Ghadisha, & Vera, 1997). Resampling also provides a similar and intuitive explanation of the finding of output interference in recall tasks (Dalezman, 1976; Dong, 1972; Roediger, 1974; Roediger & Schmidt, 1980). Across a number of experiments, it has been demonstrated that recalling a subset of the list items decreases the recall probability of the remaining items. If it is assumed that the act of recall strengthens the recalled items (as was assumed by the SAM model, Raaijmakers & Shiffrin, 1981), then the sampling probability of the recalled items will increase at the expense of the non-recalled items, producing output interference as a consequence.

Thus, while it is clear that the free recall task exhibits evidence for competition among the list items at retrieval, the success of free recall models in addressing effects of list length, list strength, and output interference comes from the usage of resampling during memory search. These effects do not necessitate inter-item similarity among the list items and a model such as ours that exhibits minimal item noise in recognition memory could be capable of exhibiting competition in a free recall task through the usage of resampling at retrieval during a free recall task.

Arguments for Item Noise Models

There have been a number of experimental findings in the recognition memory literature that have been used to argue for a larger contribution of item noise than we have estimated in our model fits, specifically the effects of manipulations of semantic similarity and decrements in performance that occur through recognition testing. Here, we discuss these arguments in detail and describe how they are in fact compatible with the results and interpretations from our modeling work.

Effects of Semantic Similarity on Recognition Memory. One line of evidence that has been used to argue for the idea that item noise plays a substantial role in recognition memory is the finding that semantic similarity among studied items impairs performance. Typically, this is accomplished in recognition experiments by increasing the number of studied items from the same semantic category, and decreases in performance with increasing category length are often observed (e.g.: Arndt & Hirshman, 1998; Criss & Shiffrin, 2004; Dewhurst & Anderson, 1999; Robinson & Roediger, 1997; Shiffrin et al., 1995). For instance, Shiffrin et al. (1995) conducted an experiment in which participants studied a large number of categories with different numbers of exemplars, where categories were defined as words associatively related to a prototype word. Robust decrements in d' were observed as category length increased from two to nine exemplars (Experiments 1 and 2, although Experiment 4 in the Appendix found no effect of category length on d' when category length was increased from one to ten exemplars). Another line of evidence concerns findings from the Deese-Roediger-McDermott (DRM: Deese, 1959; Roediger & McDermott, 1995), in which inclusion of a number of highly associated exemplars on a study list causes participants to falsely endorse a strong associate of the exemplars, a tendency which also increases with category length (Robinson & Roediger, 1997). On the surface, these results are consistent with the predictions of item noise models which predict that as more similar items are entered into memory, it becomes increasingly difficult to discriminate between studied and unstudied exemplars from the same category (Clark &

Gronlund, 1996).

While our present model fits demonstrate that the magnitude of the item noise contribution for words is quite small, we have fit data from experiments that employed lists of unrelated words. As we have previously mentioned, the magnitude of item noise is dependent on the stimulus class. Thus, one possibility is that unrelated words exhibit very low inter-item similarity whereas words that share a semantic category are sufficiently similar to generate more substantial degrees of item noise. This approach was undertaken by Johns et al. (2012) who used a holographic memory model where the item representations are high dimensional vectors generated from a large text corpus of over 30,000 documents. Specifically, each dimension in the vector reflected a particular document in the corpus and the dimension took a value of one if the word occurred in that document and a zero otherwise. The resulting vectors were quite sparse, and consequently there was virtually no overlap among the vectors of unrelated words and no effects of list length and list strength were predicted in their simulations of the model. However, similar words overlap quite substantially due to their co-occurrence in documents among the corpus, and for that reason the model was able to predict higher false alarm rates for semantically related lures in the DRM paradigm.

However, there are a number of complications from category length designs that prevent us from endorsing the view that semantically similar words exhibit high degrees of item noise. Specifically, evidence suggests that other factors complicate the interpretation of category length and DRM effects. Changes in performance across category length do not appear to be purely a consequence of having studied similar content, but also appear to reflect the usage of category labels as cues to guide retrieval.

Pure item noise models predict that as category length is increased, performance should decrease monotonically, as the higher number of similar studied items in memory makes it more difficult to discriminate between studied items and highly similar lures. However, Neely and Tse (2009) found that the change in performance is actually

non-monotonic, with performance increasing as category length increased from 1 to 2 items, decreasing to the baseline level as category length increased to eight, and further decreasing as category length increased to fourteen items. Neely and Tse (2009) suggested that the increase in performance could be achieved if additional category labels are used at retrieval, allowing memory to be more focused on items from a studied category if the category length is relatively small (such as when category length is 2 items).

When category labels get used as cues along with the studied items, the task begins to resemble an associative recognition task, as the question being asked by the participant during a recognition test is no longer “Did I see this item on the study list?" but “Is this item an exemplar of one of the categories I studied?"[9] Such a view predicts that increases in category length, which would increase the likelihood that category labels would be used as cues, should produce an increase in the ability to discriminate studied categories from unstudied categories. Dennis and Chapman (2010) found exactly this pattern. They conducted a category length experiment where category length and list length were manipulated by presenting eight categories with category lengths of one, three, and ten exemplars. Since the list length increased as the category length is increased, a pure item noise model predicts that discrimination of studied from unstudied categories should get worse with increasing length. Instead, false alarm rates to exemplars from unstudied categories decreased as category length was increased. Fits of the REM model found that the model predicted an increase in the false alarm rate to exemplars from unstudied categories. This pattern of data was well accounted for by the BCDMEM model, which exhibits no item noise, by assuming that category length increases the likelihood that category labels are used as cues in conjunction with the probe at retrieval.

Another prediction from pure item noise accounts of the category length effect is the within-category choice advantage on forced choice tests (Clark & Gronlund, 1996; Clark, 1997; Hintzman, 1988). A counter-intuitive prediction of the models is that on a 2AFC

[9]We would like to acknowledge Michael Humphreys for conceiving of this analogy.

recognition test, recognition should be superior when both choices are from the same category than when both choices are from different categories. The models make this prediction because when two similar items are presented, the memory strengths of the two items are correlated. When the difference between the two memory strengths is calculated, the covariance between the two similar items gets subtracted out, resulting in a difference distribution that has lower variance for within-category choices than for between-category choices. While Hintzman (1988) found confirmatory evidence for this prediction, Maguire, Humphreys, Dennis, and Lee (2010) noted that Hintzman compared between and within category choices as a between subjects manipulation, which may have changed the way participants approached the test. Additionally, participants were tested using booklets, which did not properly control for lag between presentations, the order of testing, etc. In the experiments of Maguire et al., categories were generated from either word association norms or taxonomically generated categories and the authors found no within-category choice advantage in all experimental conditions. They additionally confirmed the null hypothesis by using the Bayesian analysis developed by Dennis et al. (2008).

Nonetheless, a remaining question concerns why discriminability has often been found to decrease with category length. There are two such confounds that can address this regularity in category length experiments. First, when lists of semantic categories are studied, performance has been found to decrease with within-category serial position in recognition memory (Carey & Lockhart, 1973), free recall (Carey & Lockhart, 1973; Wood & Underwood, 1967), and cued recall (Jakab & Raaijmakers, 2009; Mulligan & Stone, 1999). Many investigations of category length do not control for within-category serial position, and an implication of this result is that longer categories tend to contain target items from later within-category serial positions, degrading performance. In Neely and Tse's Experiment 4, when within-category serial position was controlled there was no impairment in discriminability as category length was increased from two to fourteen exemplars. In another condition where items came from later within-category serial

positions, a category length on discriminability was found[10].

The second confound concerns a measurement issue in category length designs, namely the usage of d' to measure discriminability. A difficulty with usage of d' is its assumption of an equal variance signal detection model, which is almost uniformly violated in ROC experiments. If an equal variance model is used in the analysis, changes in the response criterion can additionally change d' (Rotello, Masson, & Verde, 2008) and for that reason, d_a is often recommended for analysis. Cho and Neely (2013) conducted a category length experiment where they employed category lengths of two, eight, and fourteen exemplars using both old/new recognition with confidence ratings and 2AFC testing. They additionally insured that for all category lengths the same number of items were tested in each category and that all items came from the same serial position in the category. While category length increased hit rates and false alarm rates and decreased d', there was no effect of category length on d_a, suggesting that a criterion shift may have been taking place. An alternative possibility is that manipulations of category length increase the mean of both the target and lure distributions to equivalent degrees. While item noise accounts make this prediction, they further predict that the variance of both distributions should increase and performance should decrease as a consequence. Nonetheless, usage of a category label as an additional cue may have the effect of shifting both distributions upward without decreasing discriminability. Additionally, Cho and Neely (2013) found no effect of category length on 2AFC recognition performance and no within-category choice advantage in the 2AFC tests, replicating the results of Maguire et al. (2010).

Another issue with category length designs concerns the production of implicit

[10]Neely and Tse (2009) proposed that attention might decrease with category length, in a manner similar to the way it has been proposed to decrease with list length. Another possibility is that decreases in performance with within-category serial position are due to prediction based learning mechanisms, whereby the encoding strength is inversely proportional to how predictable an item is given the history of learning. Prediction based learning has been used to account for primacy effects (Davelaar, 2013; Farrell & Lewandowsky, 2002; Lewandowsky & Murdock, 1989), spacing effects (Murdock, 2003), and distinctiveness effects (Farrell, 2006).

associative responses (IAR) during list presentation, which can occur when lists contain words that are strongly associated to reach other (e.g.: silk, web, legs, are associates of *spider*) rather than being members of the same taxonomic category (e.g.: spider, dog, cat are members of the category *animal*). While controlled studies have found no effect of category length using taxonomic categories (Cho & Neely, 2013; Neely & Tse, 2009), Maguire et al. (2010) found a large effect of large effect of category length on 2AFC performance for associative categories while finding no effect of category length when taxonomic categories are used. Similarly, the studies that have found DRM effects, which are possibly the biggest false memory effects found in list memory paradigms, often employ associatively related categories (Roediger & McDermott, 1995; Robinson & Roediger, 1997). The distinction between taxonomic and associative categories is relevant because during presentation of a list of strong associates, it is very likely that participants may spontaneously generate associatively related words, leading participants to falsely attribute their presentation to having occurred on the study list (Underwood, 1965).

Dennis and Humphreys (2001) proposed an IAR account of both category length and DRM effects and argued that these effects do not speak directly to similarity among the item representations. The IAR hypothesis is similar to the source monitoring account of false memory (Johnson, Hashtroudi, & Lindsay, 1993), which proposes that false memory is a consequence of the participant being unable to distinguish between events that actually occurred and imagined events. Consistent with the IAR hypothesis, Maguire et al. (2010)'s experiments using associative categories found a robust effect of category length on 2AFC performance that were accompanied by no within-category choice advantage. This result would be expected if increases in category length increased the likelihood of generating other category exemplars during study, impairing discrimination of seen from unseen category exemplars. However, if the exemplars exhibited dissimilar representations like we are hypothesizing here, no within-category choice advantage would be expected. Item noise accounts, in contrast, predict both a category length effect on discriminability and a

within-category choice advantage.

The lack of decrement in discriminability with increasing category length in controlled designs, the lack of within-category choice advantage, and the better discrimination of studied categories from unstudied categories with increasing category length are contrary to the the pure item noise account of category length effects. However, we would like to state explicitly that these results do not preclude the idea that item representations for semantically similar items exhibit more similarity to each other than unrelated words, making them more susceptible to item noise. In fact, this idea is highly plausible. Instead, these results indicate that other cognitive factors appear to be contributing to the observation of category length effects with categories constructed from word stimuli, and these factors complicate making inferences about the differing susceptibility to item noise across differring degrees of semantic similarity of the stimuli. The higher similarity among the item representations for similar words may be so negligibly small that category length would have to be manipulated to much higher numbers than in conventional experiments to observe effects consistent with the item noise account, such as the within-category choice advantage.

Additionally, we have restricted this discussion to experiments that have employed words as stimuli, as that accounts for the majority of the category length experiments that have been conducted. Our modeling results indicate that certain non-linguistic stimuli may be more susceptible to item noise than word stimuli, indicating that category length effects consistent with the item noise account should be detectable using non-linguistic stimuli. Consistent with this idea, Konkle, Brady, Alvarez, and Oliva (2010) found robust category length effects in 2AFC testing using photos of objects. Furthermore, they employed many of the same controls employed by Cho and Neely (2013)'s investigation which failed to find category length effects with words, such as always testing the same number of items for each category length and testing the same serial positions within each category.

Decrements in Performance Through Recognition Testing. One of the more recent lines of research that has been used to argue for models where item noise is the bulk of interference in recognition memory concerns the decrease in performance across test trials in a recognition memory test, which has been robustly observed in many experiments (Annis, Malmberg, Criss, & Shiffrin, 2013; Criss et al., 2011; Gillund & Shiffrin, 1984; Kim & Glanzer, 1995; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975; Peixotto, 1947; Ratcliff & Hockley, 1980; Ratcliff & Murdock, 1976; Schulman, 1974). While this finding had been known for some time, it was unclear whether the observed decrease was due to the experience of the test items or the increase in retention interval due to the passage of time. Recently, Criss et al. (2011) argued that the decrease in performance was purely due to accumulated item noise, as one of their experiments (Experiment 2) included a condition where recognition testing was delayed by a 20 minute filler task. The 20 minute delay exhibited only a relatively small decrement to performance and was much smaller than the decrement that was observed through the course of recognition testing, thus ruling out the hypothesis that the testing decrement is due to the passage of time. Converging evidence for the argument that the testing decrement reflects item noise comes from Murdock and Anderson (1975), who found that the magnitude of the testing decrement increased with the number of choices on a forced choice recognition test. If all items on each forced choice trial are added to the contents of memory, then it naturally follows that trials with more choices should exhibit greater performance decrements as a consequence of the higher item noise.

Possibly the biggest challenge to the item noise hypothesis of the testing decrement is that it is unable to explain the fact that the number of test items exhibit much larger decrements on performance than the number of study items, which typically exhibit no decrement at all under controlled conditions. For instance, Schulman (1974) compared list lengths of 25, 50, and 100 items with equated retention intervals across each list length condition and compared performance across blocks of 25 2AFC trials. For each test block,

performance was equivalent across the different list length conditions and yet there were considerable decrements in performance across the test blocks. A pure item noise account predicts poorer performance both in later test blocks and for larger list length conditions.

A plausible contender to the item-noise hypothesis is that contextual drift through the course of testing impairs performance. That is, each test trial may alter the context cue, which decreases the match to the studied items and consequently decreases recognition performance. The majority of contextual drift theories assume that items, not the passage of time, are the sources of contextual change (Mensink & Raaijmakers, 1988; Murdock, 1997; Howard & Kahana, 2002). Thus, Murdock and Anderson's (1975) observation that more choices on a forced choice trial cause greater decrements in performance can be understood as greater contextual change as a consequence of experiencing more items on each trial. While one could argue that the contextual drift explanation is ad hoc, several investigators have independently posited the idea that retrieval from episodic memory causes contextual change (Jang & Huber, 2008; Klein et al., 2007; Sahakyan & Hendricks, 2012). In Jang and Huber's investigation, they found that retrieval during episodic tasks produced greater contextual change than other forms of retrieval, which can explain why testing produced far greater decrements in recognition memory performance in Criss et al.'s (2011) investigation than that of a 20 minute retention interval.

To demonstrate that the contextual drift account can reasonably account for the testing decrement, we simulated the paradigms of Schulman (1974) and Murdock and Anderson (1975) with our model. Group-level predictions from the model for the dataset of Schulman (1974) were derived by using hyperparameters that most closely resembled the experimental parameters of the Schulman experiment. Due to the usage of high frequency word stimuli, a short study-test delay, and a two second study time, we used the context mismatch variability parameter for high frequency words, the mean context match for the three and a half minute delay, the item mismatch variability parameter for word stimuli, and the learning rate for the three second study time (which had to be multiplied by .9 to

get better resemblance to performance). All individual trials were simulated, and after each trial, two items were added to the contents of memory that corresponded to the two choices on each test trial. The mean context match for test items was set to one to reflect the greater recency for test items over study items.

The model predictions can be seen in the left panel of Figure 24. One can see that the model correctly predicts no decrement of increasing list length (the data show better performance in longer lists, although the differences were not significant), but fails to produce levels of item noise that are sufficient to produce a testing decrement as large as what is seen in the data. We simulated the contextual drift assumption by assuming that for each item on each test trial, the mean match to context for study and test items was multiplied by .9955 to reflect contextual drift caused by recognition testing. Predictions from the contextual drift assumption can be seen in the right panel of Figure 24, and one can see that it succeeds in capturing both critical aspects of the data: no effect of list length on recognition performance is predicted, while performance decreases dramatically with each block of recognition testing.

Group-level predictions for the dataset of Murdock and Anderson (1975) can be seen in Figure 25. Given that there was immediate testing in their experiment, the mean match to context parameter μ_{ss} was set to one. The study time and word frequency for the experiment was not specified, although we obtained reasonable correspondence with the data using the learning rate r from the three second study time and the context mismatch variability ρ for high frequency words. The left panel shows the predictions from the model when no contextual drift is employed and each test item is added to the contents of memory. One can see that while performance is worse as the number of choices is increased, the level of item noise is insufficient to produce any significant decrements across testing. The right panel shows the performance of the model with the additional assumption that each item on a test trial multiplies the mean match to context by .9985. One can see that the model correctly predicts a larger decrease as the number of choices on

each forced choice trial was increased.

A more recent result that has been used to argue for an item noise interpretation of the testing decrement comes from an investigation that used blocked categories at test. Malmberg et al. (2012) conducted an experiment where all studied words came from two semantic categories and included a blocked condition where the test list was divided into two blocks of 150 2AFC trials where each block tested a different semantic category. Performance decreased monotonically through the test list until the category switch point, at which point performance increased considerably and subsequently decreased over further test trials. In a control condition, all category exemplars were randomized through the test block, performance decreased monotonically through testing. Malmberg et al. (2012) argued that these results are not only consistent with item noise theories, which predict that the magnitude of interference is dependent on the similarity of the cues to the contents of memory, but are also comparable to the release from proactive interference (PI) results of Wickens and colleagues (D. D. Wickens, 1970; D. D. Wickens, Born, & Allen, 1963). In release from PI paradigms, recall of trigrams is found to decline over trials, which has been attributed to a buildup of PI from the preceding trials (Keppel & Underwood, 1962). However, when the category of the to-be tested material is suddenly shifted, such as from digits to consonants, recall improves to around the level of the first trial. This effect has been dubbed *release from PI* because it is as if the shift in the similarity of the learned material prevents memory from suffering any interference from previous trials.

While the data of Malmberg et al. are compelling, their interpretation that the release from PI effect in their data is a consequence of item noise critically assumes that the original release from PI phenomenon observed by Wickens and colleagues was also due to the similarity of the item representations. An alternative conception is that release from PI is generally due to usage of new cues at retrieval when studied categories are switched (Gardiner, Craik, & Birtwistle, 1972; Humphreys & Tehan, 1992; Tehan & Humphreys, 1995, 1996; Watkins & Watkins, 1975). Watkins and Watkins (1975) proposed a cuing

explanation of the release from PI phenomenon in their description of the cue overload principle, which states that performance degrades as the number of items associated with a cue is increased. By their view, when the category of the to-be tested material is suddenly switched, participants use a new category cue to guide their retrieval which exempts items from earlier trials to enter into the sampling set at recall.

A cuing explanation can similarly be given for the results of Malmberg et al., in that an obvious switch of the categories midway through testing may have prompted subjects to use a new category cue in conjunction with the item cue to overcome the contextual drift that had occurred throughout the test. Additionally, while there was a release from PI effect that was observed when large blocks of items were tested, no release from PI was observed when categories were shifted every five trials during testing, which is inconsistent with the item noise account of the testing decrement. One possibility is that there are costs associated with switching category cues that make it less likely when blocks are short, which is a view that is endorsed by the original authors: “It is less clear how long a block must be in order to observe a relase from output interference. There are likely costs associated with switching the contents of retrieval cues, say from emphasizing one set of item features representing category membership rather than another set." (Malmberg et al., 2012, p. 4).

Evidence for Differentiation Models of Recognition Memory. Throughout this article, we have restricted discussion of item noise to simple global matching models that predict detrimental effects of list length and list strength on recognition memory performance when the item representations bear similarity to each other. Another class of models, referred to as differentiation models, were introduced to predict the null list strength effect in recognition memory. These models include a variant of the SAM model (Shiffrin et al., 1990) along with the REM (Shiffrin & Steyvers, 1997) and SLiM (McClelland & Chappell, 1998) models. In differentiation models, the first presentation of a study list item creates a new memory trace corresponding to that item, whereas subsequent

reptitions update that memory trace. The repetitions not only make the representation more responsive to its own cue, but also makes the representation less similar to other items. Thus, strength has the functional effect of decreasing the item noise among the stored item representations with increasing strength, whereas more traditional item noise models predict increasing item noise with increased strength. Increases in list length do not induce differentiation but instead create new memory traces, and thus differentiation models predict a detrimental effect of list length on recognition memory performance.

The original motivation behind the differentiation mechanism was to predict a dissociation between list length and list strength effects. As previously mentioned, while there were many published effects showing detrimental effects of increasing list length, Dennis and colleagues (e.g.: Dennis et al., 2008) have demonstrated that these appear to be due to confounds present in list length designs that show no effect of list length for word stimuli when controlled. While non-linguistic stimuli such as fractals and faces show worse performance in longer lists (Kinnell & Dennis, 2012), the same stimuli appear to be susceptible to list strength effects as well (Osth et al., 2014). With the exception of face images, which appear to be more susceptible to manipulations of list length than list strength, in the data and modeling from our fits there appeared to be no dissociation between the effects of list length and list strength, as stimuli appeared to be affected negatively by both manipulations or they were not affected. Thus, there is not strong evidence for a dissociation between list length and list strength effects that warrants a differentiation mechanism.

Another prediction of differentiation models is the strength based mirror effect (Criss, 2006). Because increased strength makes a stored memory trace less confusable with other item representations, lures will exhibit considerably less similarity to the contents of memory when the memory traces are strong than when they are weak, making it such that the false alarm rate should reduce as memory traces are strengthened. As stated previously, this pattern is robustly observed in item recognition (Hirshman, 1995; Stretch

& Wixted, 1998b) and associative recognition (Clark & Shiffrin, 1992; Hockley & Niewiadomski, 2007; Osth & Dennis, 2014). An alternative explanation of the strength based mirror effect is that it is caused by a criterion shift. That is, higher expected memory strength for a tested study list may cause the participant to adopt a stricter response criterion for the tested items, reducing the false alarm rate.

To counter the criterion shift argument, Criss (2009, 2010) has presented evidence in support of the predictions of differentiation models by demonstrating that memory strength for lures is reduced under conditions of higher strength. Criss (2009) tested participants under pure weak and pure strong conditions and asked them to indicate their confidence on a 20 point scale. Distributions of subjective memory strength were lower for lures following a pure strong list than following a pure weak list. To counter the argument that this was the product of a criterion shift, a manipulation of target probabilities on the test list did not have any effect on the strength estimates for lures, despite the fact that it affected bias on their yes/no repsonses. In addition, Criss (2010) estimated parameters of the Ratcliff diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) and found that drift rates for lures were lower in a pure strong condition than in a pure weak condition, whereas a target probability manipulation only affected the starting point of evidence accumulation and exhibited no effect on the drift rates.

A difficulty in interpretting these results is that they do not only support differentiation models, but additionally support the overall class of likelihood ratio models. Our model, which employs a likelihood ratio transformation of the memory strengths, not only predicts lower FAR under conditions of higher strength but also predicts that the log likelihood ratio distribution for lures should also exhibit a lower mean than the corresponding distribution for lures studied in a weak study list.

A critical difference between our model and the differentiation approach lies in what generates a shift in the distribution of evidence for lures when strength is increased. In differentiation models, the strength based mirror effect is produced by encoding processes

that produce more resilient memory traces. In our model, it is the higher expected memory strength during a test that holds lures to a higher standard of evidence, producing lower likelihood ratios. Starns et al. (2010) noted that when a list of items is strengthened, both the encoding conditions and test expectations are confounded with each other. To separate the two accounts, they had participants undergo a traditional list strength paradigm with pure weak, mixed, and pure strong lists, but on the mixed lists they manipulated test expectations by only testing participants on the weak items or the strong items while informing participants about the strength composition of the test list. Differentiation models predict lower false alarm rates for stronger study lists regardless of what is expected on the test list. In contrast, false alarm rates were predicted by test expectations, as lower false alarm rates were observed in the strong test lists and higher false alarm rates were observed in the weak test lists. False alarm rates to weak test lists were nearly equivalent to false alarm rates on the pure weak study lists, and similarly false alarm rates to strong test lists were nearly equivalent to false alarm rates for the pure strong study lists.

A potential counter-argument to the results is that both differentiation and test expectations play a role in producing the strength based mirror effect. However, Starns et al. (2010) manipulated strength at two different levels. On some mixed and pure strong lists, the strong items were presented twice (strong 2X condition), whereas on others, the strong items were presented five times (strong 5X condition). False alarm rates in the pure weak tests were identical in both the strong 2X and strong 5X conditions, which can be easily explained by assuming that in both lists the lure items were held to the same expectations of memory strength. A differentiation model that also uses expected memory strength to alter the likelihood ratios predicts a lower FAR in the strong 5X condition because those memory traces are more differentiated than in the strong 2X condition. Subsequent investigations have also found that distributions of subjective memory strength (Starns, White, & Ratcliff, 2012) and drift rates for lures (Starns, Ratcliff, & White, 2012) are shifted when test expectations are manipulated following a mixed strength study list.

Thus, the available evidence suggests that differentiation is not necessary to explain either the null list strength effect or the strength based mirror effect. However, we do not mean to suggest that there is no mechanism of differentiation. Differentiation of item representations over the long term is both plausible and useful in describing how item representations evolve with experience even in cognitive domains outside of memory. For instance, McClelland et al. (1995) conducted simulations of the Rumelhart (1990) network of concept learning and demonstrated how differentiation occurs over training. The network is trained on propositions such as “A robin is a bird" and “A tree has branches." McClelland et al. (1995) discovered that during the initial training, all of the agents (robin, tree, etc.) exhibited similar hidden layer representations regardless of how similar they were to each other. As training proceeded to its conclusion, the hidden layer representations diverged and dissimilar entitites (such as robin and tree) exhibited dissimilar hidden layer representations.

One distinction between the differentiation of concepts observed by McClelland et al. and the episodic differentiation models such as SAM, REM, and SLiM, however, is that the episodic models are multiple trace models where separate copies of the item representations are stored in memory. Additionally, the episodic models posit that when a stimulus is presented in a new context, a new episodic trace is created and it is during subsequent presentations within that new context that differentiation of the episodic trace occurs (Criss, 2006, 2009). In the Rumelhart (1990) network, there is no obvious distinction between creating and differentiating representations, and differentiation operates on a larger timescale than that of a study list in a recognition memory experiments. Below, we posit that long-term differentiation may be able to explain the differences between the item noise estimates of linguistic and non-linguistic stimuli.

Differences Between Linguistic and Non-Linguistic Stimuli

Another finding in our parameter estimates was the higher estimates of item mismatch variability and item noise for fractals and faces than for words and scenes. Why would representations of faces and scenes be more susceptible to item noise? One possibility that was initially raised by Kinnell and Dennis (2012) is that fractals and faces have more overlap in their representations, making them more likely to suffer from effects of list length and list strength. But why would stimuli such as faces and fractals exhibit more overlap in their representations as opposed to words and scenes? One hypothesis is that long term experience unitizes stimulus representations to minimize within-class similarity. That is, untrained stimuli begin with overlapping representations, making them susceptible to other stored stimuli that they are similar to and thus they suffer from high degrees of item noise and background noise. However, as stimuli become unitized through training, they exhibit less similarity to other stored stimuli but still match their own previously stored representations, making them susceptible to context noise.

This hypothesis was initially proposed by Reder, Angstadt, Cary, Erickson, and Ayers (2002) to explain the non-monotonic relationship between word frequency and recognition memory performance. While low frequency words outperform high frequency words, very low frequency words exhibit worse performance than low frequency words and exhibit higher hit rates and false alarm rates (Wixted, 1992; Chalmers, Humphreys, & Dennis, 1997; Zechmeister, Curt, & Sebastian, 1978). Reder et al. (2002) conducted a training study with pseudowords and found that initial training increased hit rates and false alarm rates, but after six weeks of training, a mirror effect was evident with respect to training, with less frequently trained pseudowords exhibiting higher hit rates and lower false alarm rates. Similar results were found by Nelson and Shiffrin (2013) using Chinese characters as the stimuli. A training study using very low frequency words conducted by Chalmers and Humphreys (1998) found that definitions might facilitate unitization, in that training without definitions hurt performance on the words, whereas training with definitions

improved performance on the very low frequency words to around the level of low frequency words. Converging evidence for the unitization hypothesis could be found by observing how susceptible stimuli are to list length effects through training.

Item noise estimates for natural scene photographs closely resembled those of single words. Aside from the generally higher performance for pictorial stimuli (Brady, Konkle, Alvarez, & Oliva, 2008; Shepard, 1967; Standing, 1973), representations of scenes may resemble words due to the fact that labels can easily be applied to segments of the images. The idea that pictorial stimuli have linguistic representations was posited by Paivio (1971, 1976), who argued that pictures have “dual codes" possessing both perceptual and linguistic information. Evidence for this hypothesis comes from the finding that recognition performance for pictures is still superior to single words when the test stimuli are labels instead of the pictures themselves (Paivio, 1976; Madigan, 1983). Similarly, a mirror effect that resembles the word frequency mirror effect can be found for pictorial stimuli. Karlsen and Snodgrass (2004) found that both pictures and words rated high in familiarity exhibited lower hit rates and higher false alarm rates than those rated low in familiarity, whereas in free recall both pictures and words rated high in familiarity were better recalled.

What kind of model could explain the unitization process? Decreasing the within-class similarity for a stimulus set is very similar to the principle of differentiation employed in differentiation models such as SAM, REM, and SLiM. However, a critical distinction we would like to address is that differentiation models of episodic memory operate over the short term, whereas we argue that a differentiation-like process operates over longer time scales. Specifically, in short-term differentiation models, presentation of a familiar stimulus in a new context creates a new representation and subsequent presentations refine that newly created representation (Criss, 2006, 2009). We hypothesize that there is no distinction between creating and updating episodic item representations. Familiarization with a stimulus decreases its overlap with other stimuli of the same class, minimizing item noise and background noise, but increasing the susceptibility of the

stimulus to context noise as it becomes bound to more contexts with further experience.

The current modeling exercise gives no insight as to how the minimization of the overlap might occur. A complete model of episodic and general memory would need to describe how the item representations evolve with experience and would require long-term training data to constrain the parameters that guide the transitions (see Nelson & Shiffrin, 2013, for one such attempt). One possiblee mechanism for stimulus unitization is competitive auto-association. While our current model only describes associations between the item and context layers, within-layer connections for the items could be implemented as well (e.g.: J. A. Anderson, Silverstein, Ritz, & Jones, 1977). Distributed item representations could become more sparse over long term experience if each item layer is competitive such that only a small number of units can be active at a given time, which is achieved in neural networks by usage of the k-winner takes-all (kWTA) algorithm. Norman and O'Reilly (2003) used the k-WTA algorithm in recognition memory to show that item representations could become more sparse as a consequence of experience, indicating that such an approach shows promise for this endeavor.

On the Plausibility of the Likelihood Ratio Transformation

In our model, the mirror effect is captured via usage of a log likelihood ratio transformation of the memory strengths. Glanzer et al. (1993) argued that the usage of the likelihood ratio transformation obviates the need for a criterion shift to capture the mirror effect. Support for the usage of such a transformation comes not just from being able to capture the mirror pattern of hit and false alarm rates, but from several confirmed predictions of specific changes in the shapes of the likelihood ratio distributions in responses to experimental manipulations. Glanzer et al. (1993) demonstrated that conditions which produce better performance produce old and new distributions that are not only further from the center of the decision axis (where the log likelihood ratio is zero), but also that the variances of both the old and new distributions are higher for conditions

of better performance. Conversely, as performance degrades, both the old and new distributions converge at the center of the decision axis. Glanzer et al. (1991) referred to this prediction as *concentering*.

A consequence of concentering is that if one were to consider distributions for high and low frequency lures, or high and low frequency targets, discriminability between the two distributions corresponding to the word frequency classes improves as performance is increased. In other words, the magnitude of the word frequency effect is dependent on the level of performance. Glanzer and colleagues tested this prediction using the aforementioned 2AFC null comparison procedure of Glanzer and Bowles (1976) where two targets (null target trials) or two distracters (null distracter trials) of different word frequency classes are tested. The likelihood ratio models make the prediction in this paradigm that both the probability of choosing the LF target on null target trials ($p(LO, HO)$) and the probability of choosing the HF distracter on null distracter trials ($p(HN, LN)$) should increase under conditions of better performance. Glanzer and colleagues have confirmed this prediction for manipulations of study time (Kim & Glanzer, 1993), repetitions (Hilford et al., 1997), study-test delay (Glanzer et al., 1991), and decrements in performance with recognition testing (Kim & Glanzer, 1995). In each of these experiments, better performance usually corresponded to higher values of both $p(LO, HO)$ and $p(HN, LN)$. The latter finding is quite surprising, as the two stimuli were not studied and yet the decision between them is still influenced by the nature of the study episode, such as how much study time was allotted to the study list items.

Moreover, a recent analysis of Glanzer et al. (2009) confirmed a number of predictions from likelihood ratio models across a large number of ROC experiments. The prediction that variance of the underlying distributions increase with increases in performance was confirmed by constructing ROCs where false alarm rates from different conditions are compared (e.g.: HF FAR vs. LF FAR) and measuring the slope of the z-transformed ROC. Higher variances for better performing conditions were found across a

large number of manipulations, including word frequency, study time, repetitions, and study-test delay. Another prediction was that the overall length of the zROC points should be shorter for conditions of better performance, which was confirmed across a number of experiments and initially tested and confirmed by Stretch and Wixted (1998a).

While these investigations found evidence for the distributional predictions of likelihood ratio models, a direct test of the psychological theory behind likelihood ratio models was conducted by Wixted and colleagues. Likelihood ratio models assume that participants use information about the stimulus to make their recognition decisions. Stretch and Wixted (1998b) reasoned that if this is the case, then if participants were to study mixed lists of strong and weak items and at test were presented with cues to denote whether the item was strong or weak, participants should use this to inform their recognition decisions and exhibit a higher false alarm rate for weak cues. In contrast, the data showed that participants had equivalent false alarm rates to strong and weak cues. Morrell, Gaitan, and Wixted (2002) conducted similar experiments that yielded the same conclusions. In their experiments, participants studied a category where exemplars were repeated several times (strong category) and another category where items were only presented once (weak category). False alarm rates to weak and strong categories were equivalent. Based on these data, Wixted and colleagues argued that participants do not appear to use the strength cues on a trial-by-trial basis, casting doubt on whether participants are transforming memory evidence on the basis of expected memorability.

However, recent evidence from Starns and colleagues suggests that participants can use experimenter provided cues to inform their recognition judgments. In the aforementioned study by Starns et al. (2010), after studying a mixed list, when participants were told that they would be tested on only weak or strong items, false alarm rates very closely resembled false alarm rates from the pure weak and pure strong tests. Further tests of the Stretch and Wixted (1998b) procedure have also found that participants can adjust expectations to the colors that denote different levels of strength.

Hicks and Starns (2014) found that participants exhibited different false alarm rates to different color strength cues when strong and weak items were tested in separate blocks of 40 items. Starns and Olchowski (2014) found a similar compliance with colored strength cues when the weak and strong items required different response keys using randomized presentation of the strength cues.

Similarly, other studies have found different false alarm rates for conditions outside of the color cue procedure. Singer and Wixted (2006) tested categories from different retention intervals and found different false alarm rates for immediate and delayed categories when the two categories were studied two days apart. Furthermore, an ROC analysis that compared the false alarm rates from the immediate and delayed categories revealed higher variability for the immediate categories, a result which is consistent with the predictions of likelihood ratio models. Singer (2009) found different false alarm rates for strong and weak categories when a pleasantness ratings task was used at encoding, contrary to the results of Morrell et al. (2002). While these results show a willingness to use experimenter provided cues to inform recognition decisions appears to support likelihood ratio models, why do these data provide support for while the investigations of Stretch and Wixted (1998b) and Morrell et al. (2002) did not? One possibility is that participants regularly rely on likelihood ratios to make recognition decisions, but have difficulty in mapping their expected memorabilities to novel experiment provided cues such as color, which would not be expected to be predictive of memory strength a priori. Procedures such as blocking and making different responses may facilitate usage of the strength cues.

Other Sources of Variability in Recognition Memory

One source of variability that was not included in the model fit was trial-to-trial variability in criterion placement. A number of criterion noise models of signal detection have been developed that include such variability (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008; Wickelgren, 1968). However, one of the critical challenges in

incorporating criterion variability into a model fit is identifying its contribution from the variability in memory strength. Benjamin et al. (2009) proposed that such measurement can be done using an ensemble recognition paradigm in which participants give yes/no decisions not to one stimulus, but to ensembles of old and new words that vary in the number of words contained in the ensemble, with the logic being that ensemble size constrains memory strength variability but does not affect criterion variability. However, the approach of Benjamin et al. (2009) remains somewhat controversial, as Kellen, Klauer, and Singmann (2012) have argued that Benjamin et al. (2009) severely over-estimated criterion noise in their dataset. They demonstrated in their model fit that if decision criteria are allowed to shift across the ensemble sizes, criterion variability estimates decrease dramatically.

Another approach that avoids the usage of new paradigms involves usage of response time models to quantify decision noise. The majority of current sequential sampling models of response time employ trial-to-trial variability in the starting point of evidence accumulation (S. D. Brown & Heathcote, 2008; Ratcliff et al., 1999; Usher & McClelland, 2001), a source of decision noise which has been described as analogous to criterion variability in a signal detection framework (Benjamin, 2013; Benjamin et al., 2009). While sequential sampling models are quite successful in their ability to measure aspects of the decision-making process such as response caution, bias, and the strength of the evidence, they are not able to specify the contributions of encoding and interference that are contributing to the evidence used in the decision. One possible extension of our model is to use a back-end sequential sampling model to produce decisions, allowing the model to not only make response time predictions but also to estimate the contribution of decision noise. We were not able to undertake such an approach in the current investigation because much larger numbers of correct and error responses are needed in each response category than are contained in our present datasets to properly estimate the response time distributions (Ratcliff & McKoon, 2008).

Conclusion

Our fits of a global matching model which parameterizes the matches and mismatches to item and context demonstrated that the bulk of interference comes from experiences prior to the list-learning episode (context noise and background noise), with confusable stimuli such as fractals and faces exhibiting at most small contributions of item noise. While these parameter estimates may seem counter-intuitive, they appear to be quite consistent with a wide variety of findings in the recognition memory literature as well as theories in cognitive neuroscience that advocate sparse distributed item representations.

Additionally, the model was able to fit quite well to a variety of manipulations of stimulus class, strength, list length, list strength, study-test delay, and word frequency. Several of these variables have been considered challenging to the first generation of global matching models, while the results of our modeling work suggest that the initial global matching models are quite capable of addressing these results by parameterizing the similarities between the representations. An additional advantage to parameterizing similarities instead of vectors is that it obviates the need for a vector size parameter, which affects the identifiability of the model parameters (Montenegro et al., 2011; Myung et al., 2007). Small effects of list length and list strength are well accommodated by low values of item mismatch, unequal variance between targets and lures can be explained with item and context match variability, and mirror effects are capably explained by a likelihood ratio transformation of memory strengths. The global matching model we present is both simple and tractable, and shows promise in being extended to other memory tasks.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., ... Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–1030.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model. *Psychological Review*, *84*(5), 413–451.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123.

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

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistical Surveys*, *4*, 40–79.

Arndt, J., & Hirshman, E. (1998). True and False Recognition in MINERVA2: Explanations from a Global Matching Perspective. *Journal of Memory and Language*, *39*, 371–391.

Benjamin, A. S. (2013). Where is the criterion noise in recognition? (Almost) everyplace you look: Comment on Kellen, Klauer, and Singmann (2012). *Psychological Review*, *120*(3), 720–726.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural netowrk model. *Psychological Review*, *113*, 201–233.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory and Cognition*, *11*, 307–315.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has

a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127–181.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology*, *29*(3), 461–473.

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

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of repetition on associative recognition in young and older adults: Item and associative strengthening. *Psychology and Aging*, *26*(1), 111–126.

Buchler, N. G., Light, L. L., & Reder, L. M. (2008). Memory for items and associations: Distinct representations and processes in associative recognition. *Journal of Memory and Language*, *59*, 183–199.

Carey, S. T., & Lockhart, R. S. (1973). Encoding differences in recognition and recall. *Memory & Cognition*, *1*(3), 297–300.

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

Chalmers, K. A., & Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology-Learning Memory and Cognition*, *24*(3), 610–632.

Chalmers, K. A., Humphreys, M. S., & Dennis, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. *Memory & Cognition*, *25*(6), 780–784.

Chappell, M., & Humphreys, M. S. (1994). An auto-associative neural network for sparse

representations: Analysis and application to models of recognition and cued recall. *Psychological Review*, *101*, 103–128.

Cho, K. W., & Neely, J. H. (2013). Null category-length and target-lure relatedness effects in episodic recognition: A constraint on item-noise interference models. *The Quarterly Journal of Experimental Psychology*, *66*(7), 1331–1355.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 232–238.

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

Clark, S. E., & Hori, A. (1995). List length and overlap effects in forced-choice associative recognition. *Memory & Cognition*, *23*(4), 456–461.

Clark, S. E., & Shiffrin, R. M. (1992). Cuing Effects and Associative Information in Recognition Memory. *Memory & Cognition*, *20*(5), 580–598.

Cleary, A. M., Curran, T., & Greene, R. L. (2001). Memory for detail in item versus associative recognition. *Memory & Cognition*, *29*(3), 413–423.

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

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

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology*, *59*, 297–319.

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

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

Criss, A. H., & Shiffrin, R. M. (2004). Context-noise and item-noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, *111*, 800–807.

Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 597–608.

Davelaar, E. J. (2007). Sequential retreival and inhibition of parallel (re)activated representations: A neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, *15*(1), 51–71.

Davelaar, E. J. (2013). A novelty-induced change in episodic (NICE) context account of primacy effects in free recall. *Psychology*, *4*(9), 695–703.

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

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*(4), 710–721.

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 18–33.

Deese, J. (1959). On the prediction of the occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language*, *63*, 416–424.

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

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian Analysis of Recognition Memory:

The Case of the List-Length Effect. *Journal of Memory and Language*, *59*, 361–376.

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & Cognition*, *27*, 664–673.

Dong, T. (1972). Cued partial recall of categorized words. *Journal of Experimental Psychology*, *93*(1), 123–129.

Dosher, B. A., & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General*, *118*(2), 191–211.

Egan, J. P. (1958). *Signal detection theory and ROC analysis* (Tech. Rep.). Hearing and Communication Laboratory.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, *89*(6), 627–661.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403–408.

Farrell, S. (2006). Mixed-list phonological similarity effects in delayed serial recall. *Journal of Memory and Language*, *55*, 587–600.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin and Review*, *9*, 59–79.

Gallo, D. A., Sullivan, A. L., Daffner, K. R., Schacter, D. L., & Budson, A. E. (2004). Associative recognition in Alzheimer's disease: Evidence for impaired recall-to-reject. *Neuropsychology*, *18*(3), 556–563.

Gardiner, J. M., Craik, F. I. M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior*, *11*, 778–783.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall.

Psychological Review, *91*(1), 1–67.

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

Glanzer, M., & Adams, J. K. (1990). The Mirror Effect in Recognition Memory: Data and Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5–16.

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 81–93.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human leanring and memory*, *2*(1), 21–31.

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

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500–513.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Huntington, NY: Krieger.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1335–1369.

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

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*, 73–137.

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *42*, 742–754.

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition*, *25*(5), 593–605.

Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review*, *95*(4), 528–551.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 302–313.

Hockley, W. E., & Cristi, C. (1996a). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*(2), 202–216.

Hockley, W. E., & Cristi, C. (1996b). Tests of the separate retrieval of item and associative information using a frequency-judgment task. *Memory & Cognition*, *24*(6), 796–811.

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition*, *35*(4), 679–688.

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

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different Ways to Cue a Coherent Memory System - a Theory for Episodic, Semantic, and Procedural Tasks. *Psychological Review*, *96*(2), 208–233.

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

Humphreys, M. S., & Tehan, G. (1992). A simultaneous examination of recency and cuing

effects. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 143–159). Erlbaum.

Jakab, E., & Raaijmakers, J. G. W. (2009). The role of item strength in retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 607–617.

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

Jang, Y., Mickes, L., & Wixted, J. T. (2012). Three tests and three corrections: A comment on Koen and Yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 513–523.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*(4), 486–518.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Jordan, M. I. (2004). Graphical Models. *Statistical Science*, *19*, 140–155.

Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(5), 933–953.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*, 1163–1177.

Karlsen, P. J., & Snodgrass, J. G. (2004). The word-frequency paradox for recall/recognition occurs for pictures. *Psychological Research*, *68*, 271–276.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457–479.

Kelley, R., & Wixted, J. T. (2001). On the Nature of Associative Information in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 701–722.

Keppel, G., & Underwood, B. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior*, *1*, 153–161.

Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 638–652.

Kim, K., & Glanzer, M. (1995). Intralist interference in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1096–1107.

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

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

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

Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recolleciton and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1536–1542.

Koen, J. D., & Yonelinas, A. P. (2013). Still no evidence for the encoding variability hypothesis: A reply to Jang, Mickes, and Wixted (2012) and Starns, Rotello, and Ratcliff (2012). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 304–312.

Konkle, T., Brady, T. F., Alvarez, G., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

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

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities, and rules in category representation: An example of Hierarchical Bayesian Analysis. *Cognitive Science*, *32*(8), 1403–1424.

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

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*, 25–57.

Light, L. L., Patterson, M. M., Chung, C., & Healy, M. R. (2004). Effects of repetition and response deadline on associative recognition in young and older adults. *Memory & Cognition*, *32*(7), 1182–1193.

Madigan, S. (1983). Picture memory. In J. C. Yuillie (Ed.), *Imagery, memory, and cognition: Essays in honor of Allan Paivio* (pp. 65–89). Erlbaum.

Maguire, A. M., Humphreys, M. S., Dennis, S., & Lee, M. D. (2010). Global similarity accounts of embedded-category designs: Tests of the global matching models. *Journal of Memory and Language*, *63*(2), 131–148.

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

Science, *23*(2), 115–119.

Malmberg, K. J., & Shiffrin, R. M. (2005). The “one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336.

Malmberg, K. J., & Xu, J. (2007). On the flexibility and the fallibility of associative memory. *Memory & Cognition*, *35*(3), 545–556.

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

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

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

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

Mickes, L., Johnson, E. M., & Wixted, J. T. (2010). Continuous recollection versus unitized familiarity in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 843–863.

Montenegro, M., Myung, J. I., & Pitt, M. A. (2011). *REM integral expressions.* (Unpublished manuscript)

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection based moels of recognition memory. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *28*(6), 1095–1110.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–494.

Mulligan, N., & Stone, M. (1999). Attention and conceptual priming: Limits on the effects of divided attention in the category-exemplar production task. *Journal of Memory and Language*, *41*, 253–280.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626.

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

Murdock, B. B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review*, *10*(3), 570–588.

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

Murdock, B. B., & Kahana, M. J. (1993a). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 689–697.

Murdock, B. B., & Kahana, M. J. (1993b). List-Strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1450–1453.

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

Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

17(5), 855–874.

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition*, *19*(2), 119–130.

Myung, J. I., Montenegro, M., & Pitt, M. A. (2007). Analytic expressions for the BCDMEM model of recognition memory. *Journal of Mathematical Psychology*, *51*, 198–204.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *2000*(26), 1170–1187.

Neely, J. H., & Tse, C.-S. (2009). Category length produces an inverted-U discriminability function in episodic recognition memory. *The Quarterly Journal of Experimental Psychology*, *62*(6), 1141–1172.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.

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

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary leanring systems approach. *Psychological Review*, *110*(4), 611–646.

Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin and Review*, *15*(1), 36–43.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception*

and Performance, *17*(1), 3–27.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311–345.

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

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *Quarterly Journal of Experimental Psychology*, *67*(9), 1826–1841.

Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart, & Winston.

Paivio, A. (1976). Imagery in recall and recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 103–129). Wiley.

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

Pike, R. (1984). Comparison of Convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*(3), 281–294.

Pitt, M. A., & Myung, J. I. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.*

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

Pooley, J. P., Lee, M. D., & Shankle, W. R. (2011). Understanding memory impairment with memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, 47–56.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual process models of recognition memory. *Journal of Mathematical Psychology*, *55*(1), 36–46.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224–232.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review*, *88*(2), 93–134.

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

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect : I. Data and discussion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *16*(2), 163–178.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. In R. S. Nickerson (Ed.), *Attention and Performance VIII.* Erlbaum.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory: Receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763–785.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval Processes in Recognition Memory. *Psychological Review*, *83*(3), 190–214.

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

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

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of

reaction time. *Psychological Review*, *106*(2), 261–300.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 138–152.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*(3), 365–372.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 231–237.

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition*, *2*(2), 261–269.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not present in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 803–814.

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *6*(1), 91–105.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389–401.

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

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). Academic Press.

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

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

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

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

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156–163.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 267–287.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

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

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

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition*, *37*(7), 976–984.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition*, *34*, 125–137.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2).

Starns, J. J., & Olchowski, J. E. (2014). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*.

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in

recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language*, *70*, 36–52.

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

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

Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Mixing strong and weak targets provides no evidence against the unequal-variance explanation of zROC slope: A comment on Koen and Yonelinas (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 793–801.

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

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition*, *40*(8), 1189–1199.

Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1397–1410.

Stretch, V., & Wixted, J. T. (1998b). On the differences between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379–1396.

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

Tehan, G., & Humphreys, M. S. (1995). Transient phonemic codes and immunity to proactive interference. *Memory & Cognition*, *23*(2), 181–191.

Tehan, G., & Humphreys, M. S. (1996). Cuing effects in short-term recall. *Memory & Cognition*, *24*(6), 719–732.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, *2*(2), 189–200.

Tulving, E., & Hastie, R. (1972). Inhibition Effects of Intralist Repetition in Free Recall. *Journal of Experimental Psychology*, *92*(3), 297–304.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

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

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

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology*, *70*, 122–129.

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

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

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

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

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition*, *30*(6), 885–892.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *104*(4), 442–452.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, *5*, 102–122.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*, *3*, 316–347.

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*(1), 1–15.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440–445.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 681–690.

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

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure and mixed-strength Lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 523–538.

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

Wood, G., & Underwood, B. J. (1967). Implicit responses and conceptual similarity.

Journal of Verbal Learning & Verbal Behavior(6), 1–10.

Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal and nonverbal pair strength on associative recognition. *Memory & Cognition*, *35*(3), 526–544.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341–1354.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the List-Strength Effect in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 345–355.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society*, *11*, 371–373.

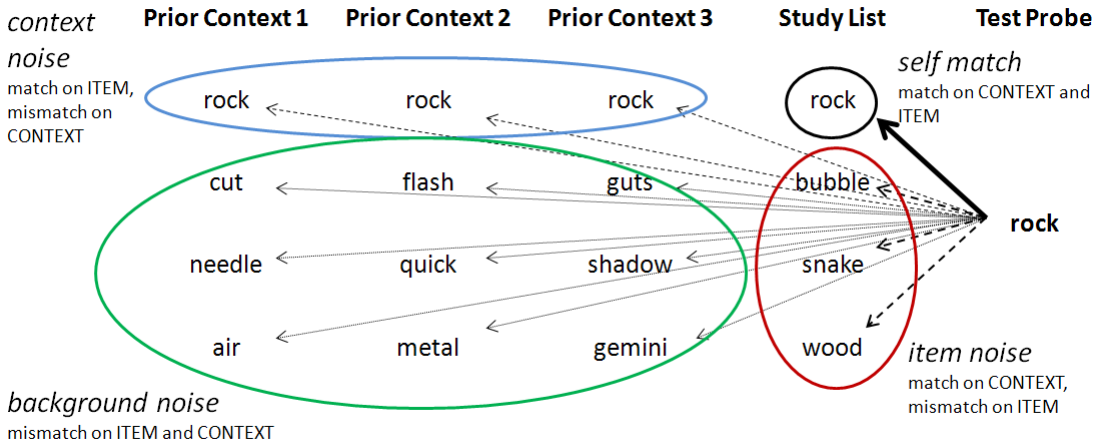

Figure 1. Diagram of the different sources of interference on retrieval within a global matching model. Depicted is a simplification the contents of memory, which include memories formed during the study list episode and memories from prior contexts. The probe cue and the study context are matched against the contents of memory simultaneously. The self match refers to an exact match on item and context information, in that the cues are matched against a representation of the cue item formed during the study list episode. Item noise refers to a match on context information but a mismatch on context information, in that the items are different from the probe cue but the memories were formed during the list episode. Context noise refers to a match on item information but a mismatch on context information, in that the memories are of the cue item but were formed prior to the list episode. Background noise refers to a mismatch on both item and context information. The magnitudes of each source of interference depend on the similarities between the matches and mismatches on item and context information.

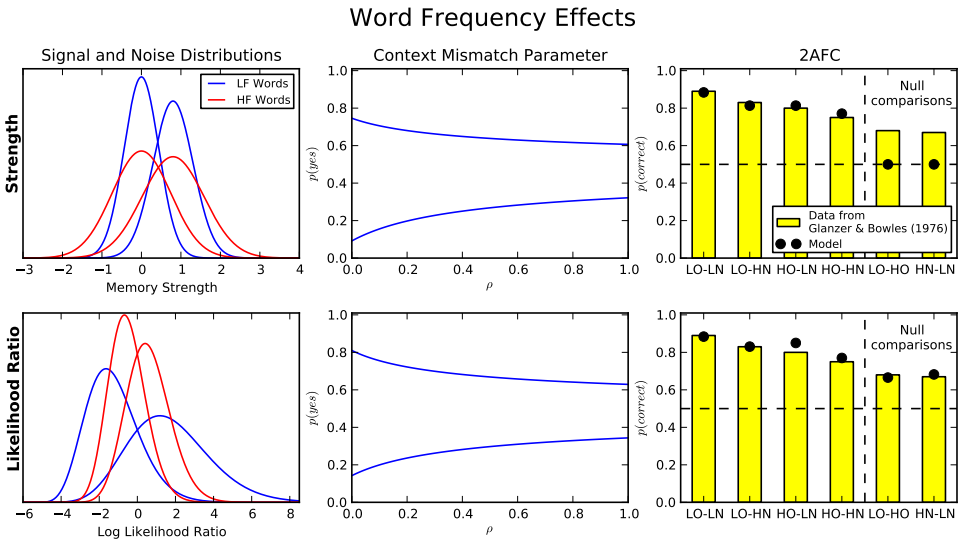

Figure 2. Graph displaying how changes in the context mismatch variability can account for the word frequency mirror effect using both the raw memory strengths (top row) and after the log likelihood ratio transformation (bottom). In the left column are signal and noise distributions for low ($\rho = .025$) and high ($\rho = .4$) values of the context mismatch variability paramter, which correspond to low and high frequency words. In the middle column are predicted hit and false alarm rates for values of the context mismatch variability parameter ranging from .025 to 1.0. In the right column are the data from the 2AFC paradigm employed by Glanzer and Bowles (1976) along with model predictions (LF: $\rho = .05$, HF: $\rho = .4$). All other model parameters are as follows: list length of 30 items, $r_{item} = .4$, $\mu_{tt} = 1$, $\mu_{ss} = 1$, $\sigma^2_{tt} = .025$, $\sigma^2_{ss} = .075$, $\sigma^2_{ti} = .002$, $\beta_{item} = .1$.

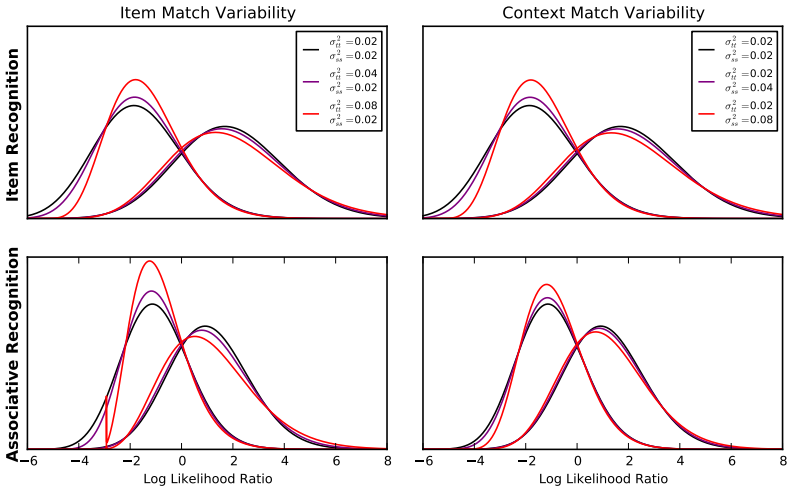

Figure 3. Graph displaying the effects of changes on the item match and context match variability parameters (σ^2_{tt} and σ^2_{ss}). Both parameters were initially set to .001. As both parameters are increased, the variability of the target distribution increases more than the variability of the lure distribution. Other parameters of the model were set as follows: list length = 20, $r_{item} = 1$, $r_{assoc} = 1$, $\mu_{tt} = 1$, $\mu_{ss} = 1$, $\sigma^2_{ti} = .002$, $\rho = .1$, $\beta_{item} = .15$, $\beta_{assoc} = .45$.

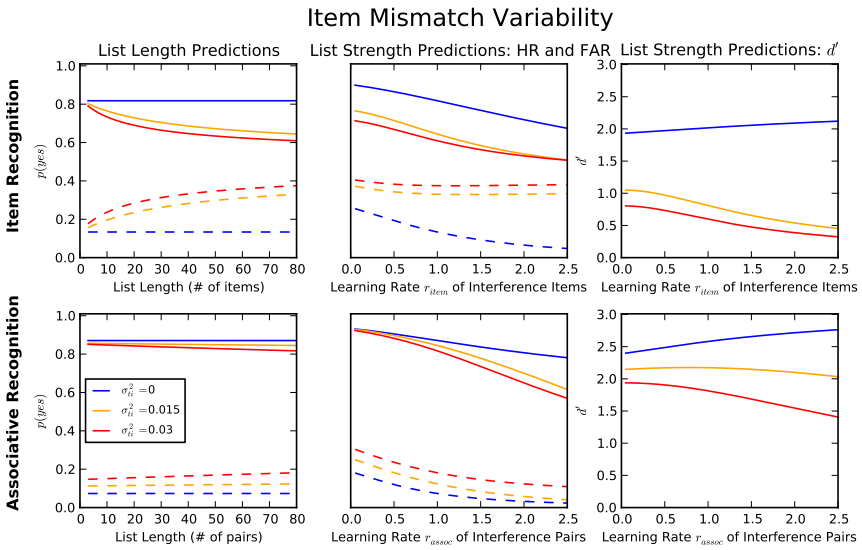

Figure 4. Graph displaying demonstrations of the effects of the item mismatch variability parameter on both item recognition (top row) and associative recognition (bottom row), which controls the amount of item noise in the model. Depicted are simulations of a list length paradigm (left) and a list strength paradigm (middle column: HR and FAR predictions, right column: d' predictions). In the list length paradigm, the number of study list items was manipulated between 1 and 80. In the list strength paradigm, 30 items were studied, half were baseline pairs studied with learning rate $r_{item} = 1.0$ and the other half were interference items studied with r_{item} ranging between .05 and 2.5. The other parameters of the model were set as follows: $r_{item} = 1$, $r_{assoc} = 1$, $\mu_{tt} = 1$, $\mu_{ss} = 1$, $\sigma^2_{tt} = .025$, $\sigma^2_{ss} = .075$, $\rho = .1$, $\beta_{item} = .1$, and $\beta_{assoc} = .1$

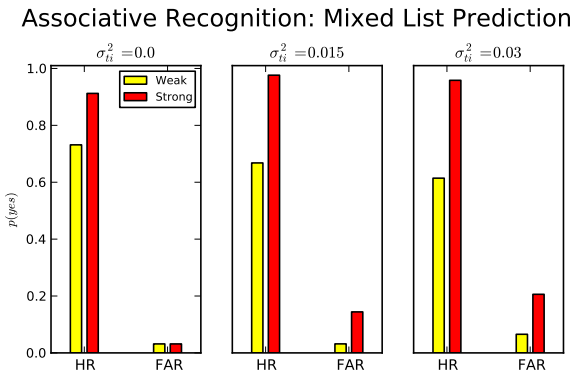

Figure 5. Graph displaying demonstrations of the effects of the item mismatch variability parameter ($\sigma_{ti}^2 = 0, .015, .03$) on a mixed list of weak and strong pairs in associative recognition. Parameters were as follows: 15 weak pairs encoded with $r_{assoc} = 1.0$, 15 strong pairs encoded with $r_{assoc} = 2.5$, $\mu_{tt} = 1$, $\mu_{ss} = 1$, $\sigma_{tt}^2 = .025$, $\sigma_{ss}^2 = .075$, and $\beta_{assoc} = .1$

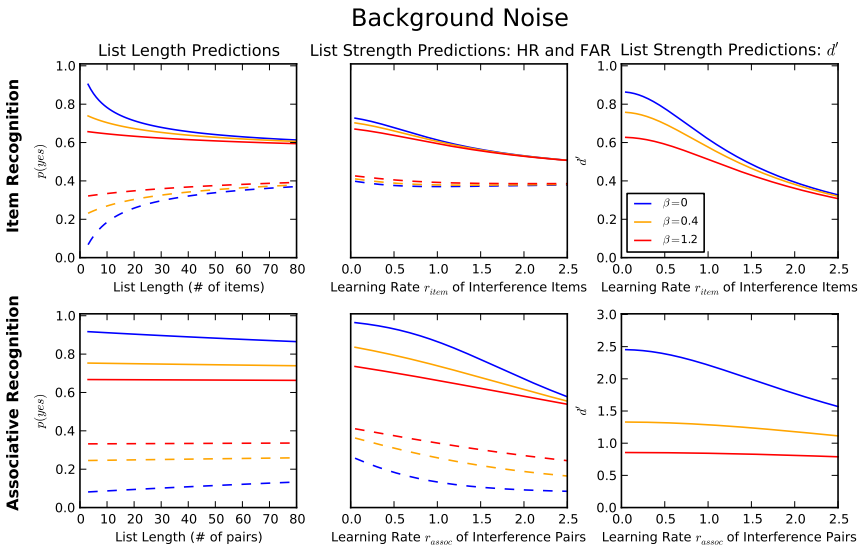

Figure 6. Graph displaying demonstrations of the effects of background noise both item recognition (top row) and associative recognition (bottom row). Background noise (β_{item} and β_{assoc}) was varied between 0, .4, and 1.2. Depicted are simulations of a list length paradigm (left) and a list strength paradigm (middle column: HR and FAR predictions, right column: d' predictions). In the list length paradigm, the number of study list items was manipulated between 1 and 80. In the list strength paradigm, 30 items were studied, half were baseline pairs studied with learning rate $r_{item} = 1.0$ and the other half were interference items studied with r_{item} ranging between .05 and 2.5. The other parameters of the model were set as follows: $r_{item} = 1$, $\mu_{tt} = 1$, $\mu_{ss} = 1$, $\sigma^2_{tt} = .025$, $\sigma^2_{ss} = .075$, $\sigma^2_{ti} = .03$, $\rho = .001$.

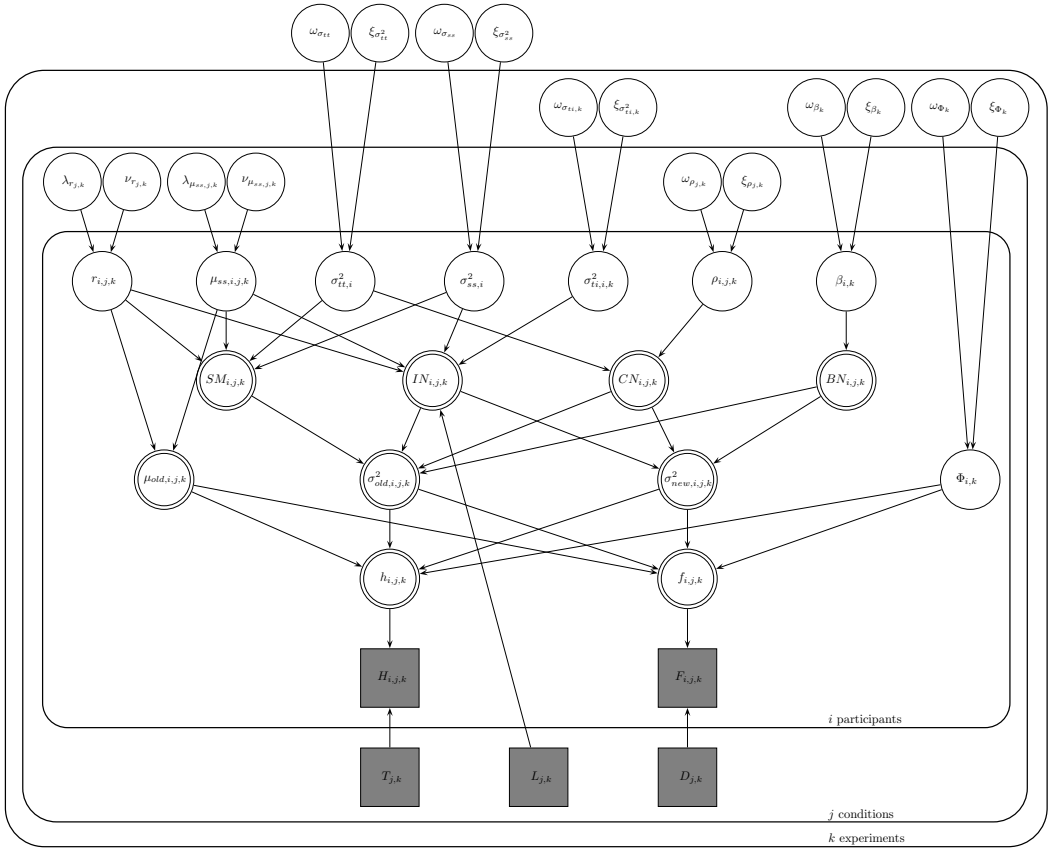

Figure 7. General graphical model representation of the hierarchical Bayesian fit of the recognition memory model. Description of the entire set of hyperparameters can be found in Table 3. Note: SM = self match, IN = item noise, CN = context noise, BN = background noise.

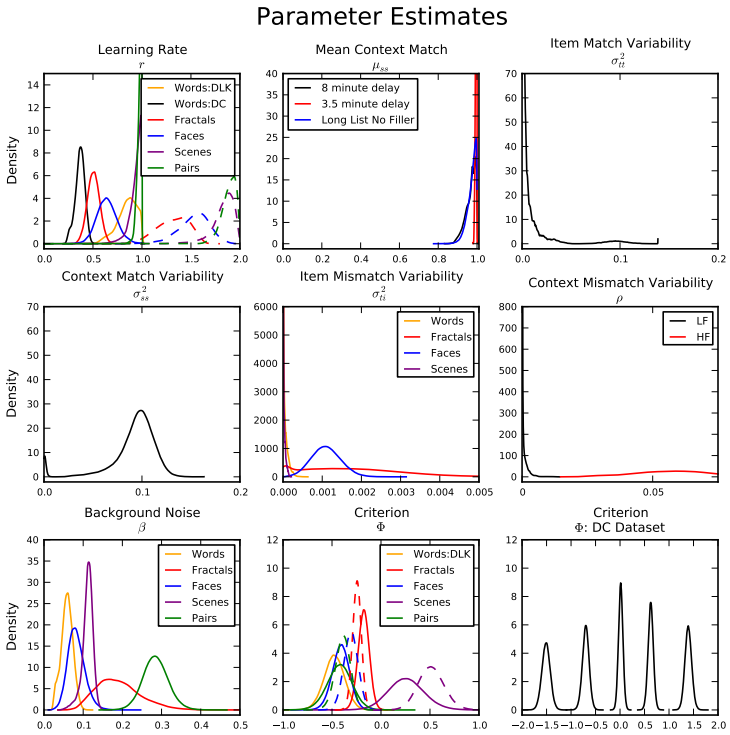

Figure 8. Posterior distributions for all hyperparameters in the omnibus fit. Depicted parameters are the learning rate r (learning rates for strong items/pairs are the sum of the weak and strong learning samples and are indicated by dashed lines), the mean context match μ_{ss}, the variance in the item match σ^2_{tt}, the variance in the context match σ^2_{ss}, the item mismatch variance σ^2_{ti}, the context mismatch variance ρ, the background noise β, along with the decision criteria Φ. Note: DLK = Dennis, Lee, and Kinnell (2008) dataset, DC = DeCarlo (2007) dataset.

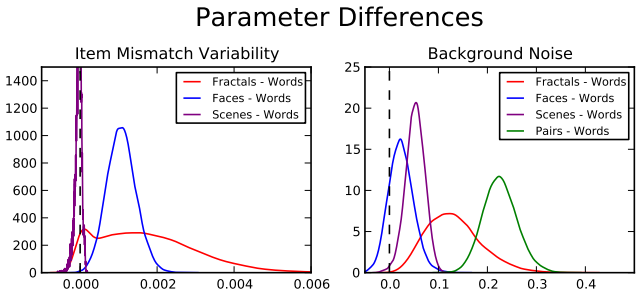

Figure 9. Density estimates of the differences between the means of the hyperparameters between words and other stimulus classes for both the item mismatch variability parameter σ_{ti}^2 (left) and the background noise parameter β (right).

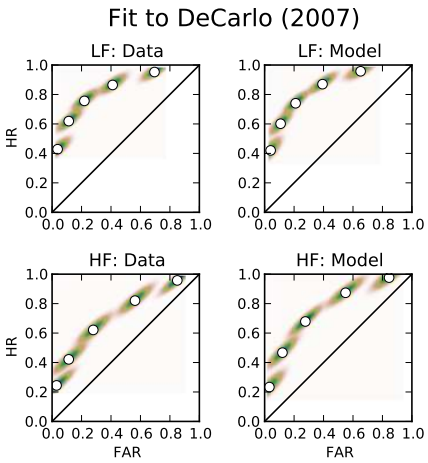

Figure 10. Density estimates of the ROC function for both the data of DeCarlo (2007, left) and model (right) for low (top) and high (bottom) frequency words. The circles indicate the median of the hit and false alarm rate posterior distributions for each ROC point.

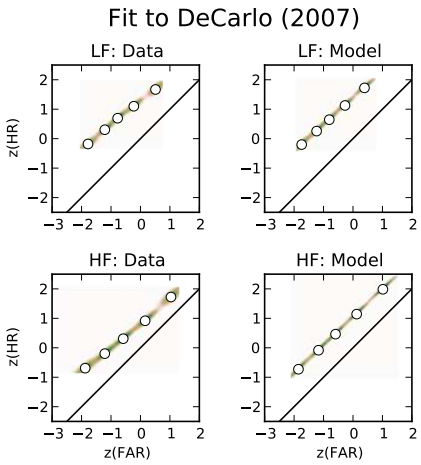

Figure 11. Density estimates of the zROC function for both the data of DeCarlo (2007, left) and model (right) for low (top) and high (bottom) frequency words. The circles indicate the median of the z-transformed hit and false alarm rate posterior distributions for each ROC point.

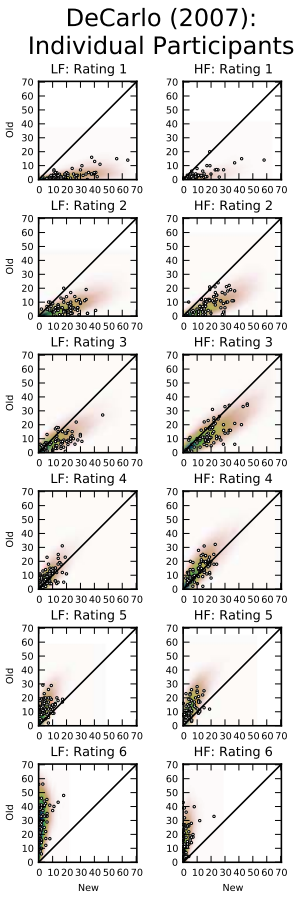

Figure 12. Individual participant responses for each confidence rating and word frequency class for the data of DeCarlo (2007). Predicted confidence counts from the model are shown as the colored density estimates.

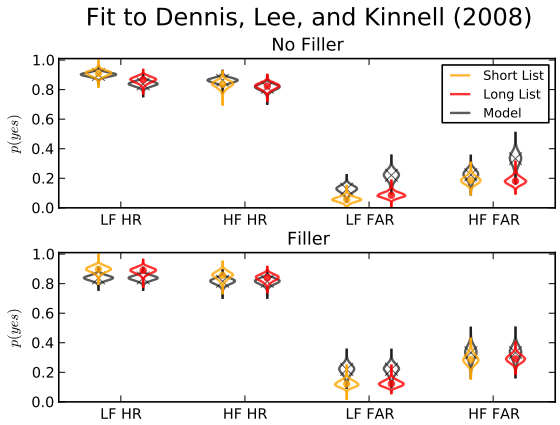

Figure 13. Fit to the data of Dennis et al. (2008) for all conditions. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

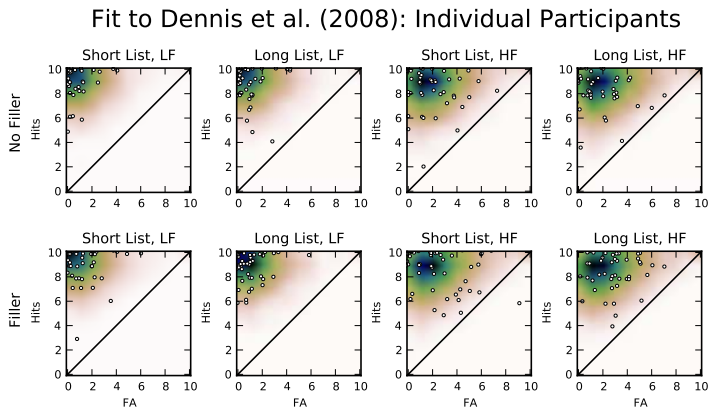

Figure 14. Individual participant hit and false alarm counts from the Dennis et al. (2008) dataset for all conditions. Predicted hit and false alarm counts from the model are shown as the colored density estimates.

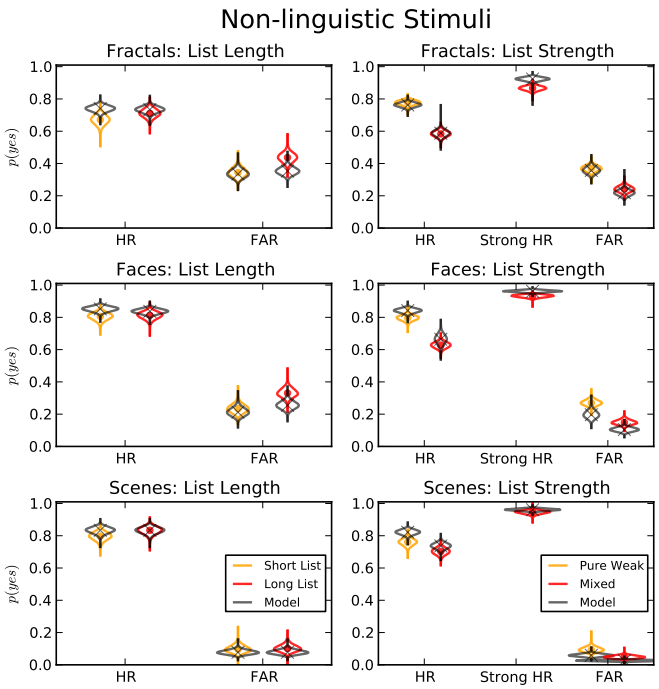

Figure 15. Hit and false alarm rates for the data and the model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth et al. (2014, right). Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

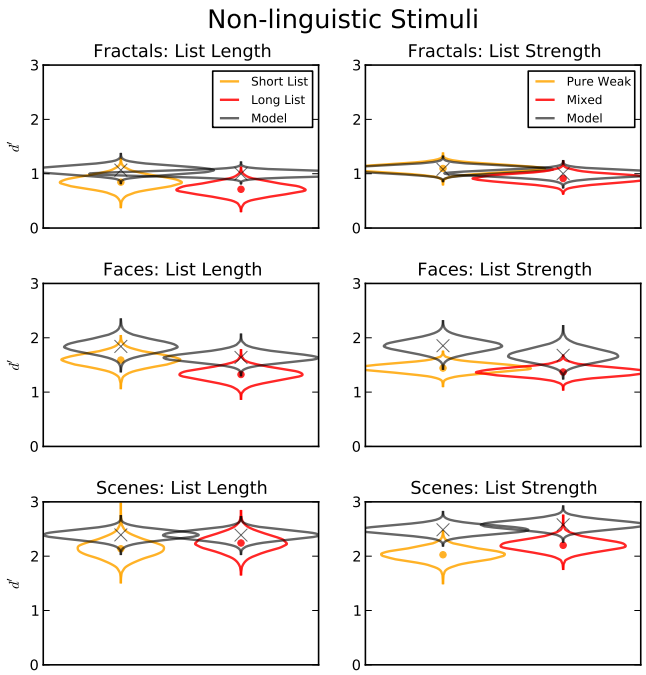

Figure 16. Hit and false alarm rates for the data and the model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth et al. (2014, right) that employ non-linguistic stimuli. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

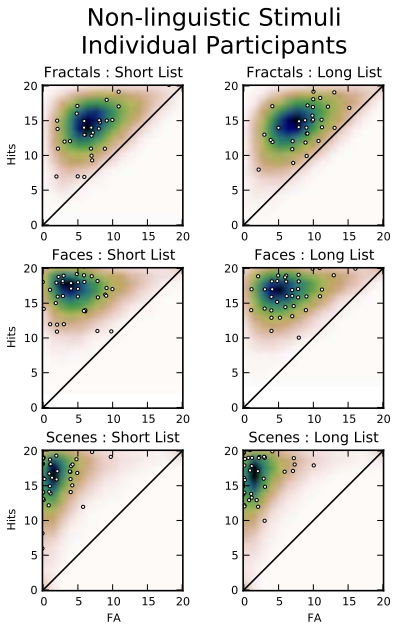

Figure 17. Individual participant hit and false alarm counts for the short (left) and long (right) list conditions for fractals (top), faces (middle), and scenes (bottom). Predicted hit and false alarm counts from the model are shown as the colored density estimates.

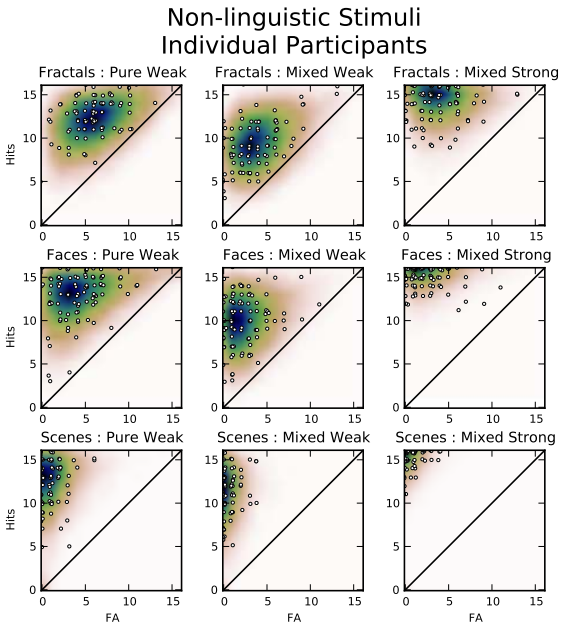

Figure 18. Individual participant hit and false alarm counts for the pure weak (left) and mixed weak (middle) and mixed strong (right) conditions for fractals (top), faces (middle), and scenes (bottom). Predicted hit and false alarm counts from the model are shown as the colored density estimates. False alarms are the same for the mixed weak and mixed strong conditions while the hits to 4X presented items are in the mixed strong plot.

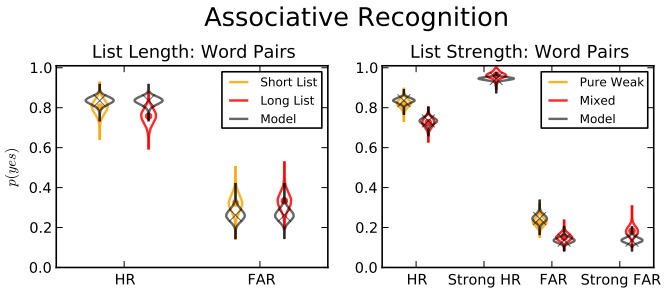

Figure 19. Hit and false alarm rates for the data and the model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth and Dennis (2014, right) that employ word pairs in an associative recognition task. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

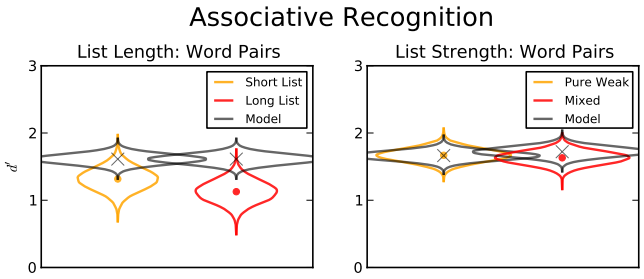

Figure 20. Hit and false alarm rates for the data and the model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth and Dennis (2014, right) that employ word pairs in an associative recognition task. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

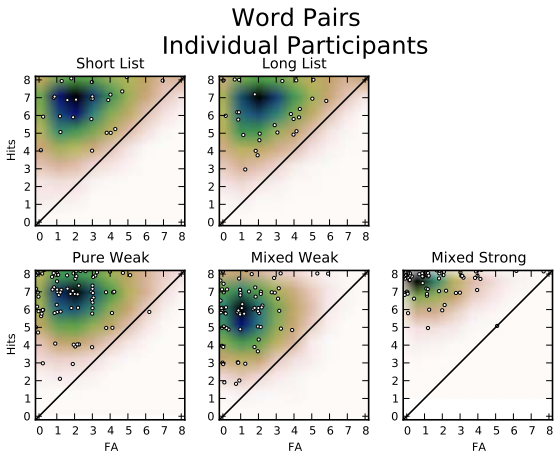

Figure 21. Individual participant hit and false alarm counts for list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth and Dennis (2014, right) that employ word pairs in an associative recognition task. Predicted hit and false alarm counts from the model are shown as the colored density estimates. For the mixed strong condition, hits and false alarm counts are from the strong intact and strong rearranged pairs.

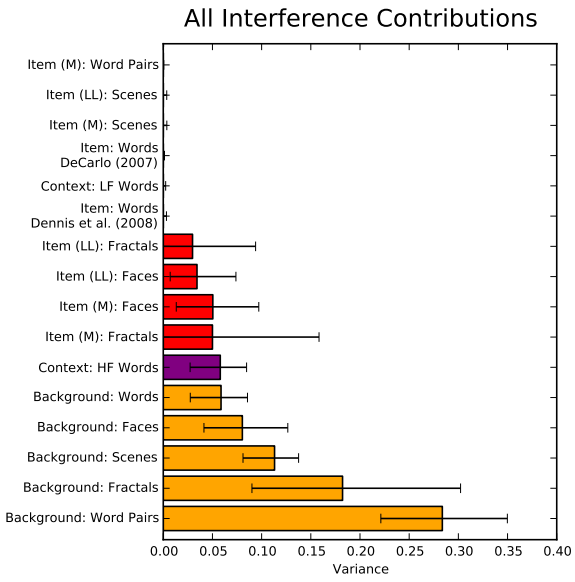

Figure 22. Density estimates for item noise, context noise, and background noise for all datasets. Notes: LF = low frequency, HF = high frequency, LL = list length, LS = list strength.

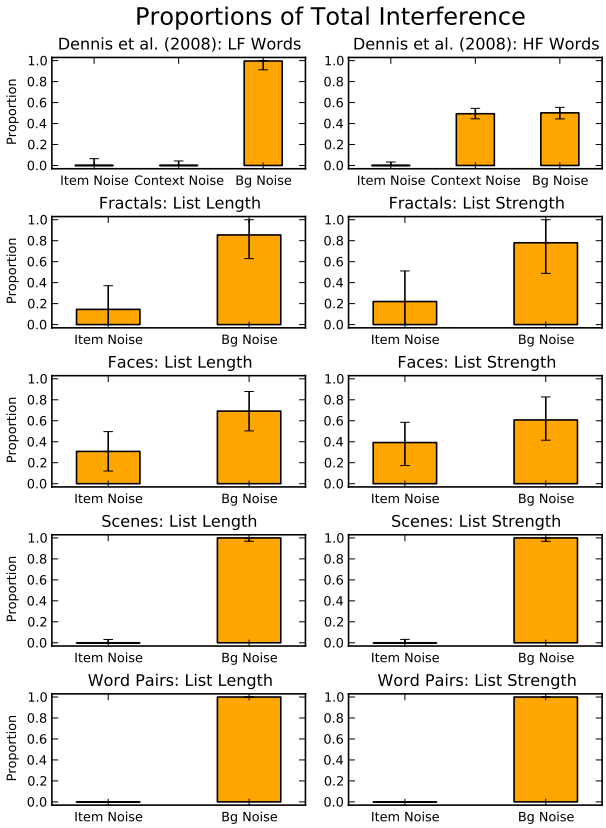

Figure 23. Median proportions of total interference for item noise, context noise, and background noise (bg. noise) for the datasets of Dennis et al. (2008), Kinnell and Dennis (2012), Osth et al. (2014), and Osth and Dennis (2014). Error bars represent the 95% highest density interval (HDI). Notes: LF = low frequency, HF = high frequency.

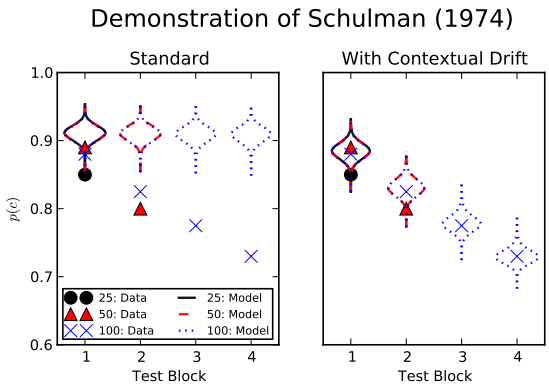

Figure 24. Group level predictions for the Schulman (1974) paradigm using parameters derived from the model fit, where on each individual trial two items are added to the contents of memory. The left panel shows the model predictions without parameter modification, while the right panel shows the model's predictions where each item on a test trial multiplies the mean context match by .9955. Due to an inavailability of the original data from the Schulman (1974) dataset, the data means were approximated from the graphs in the original article.

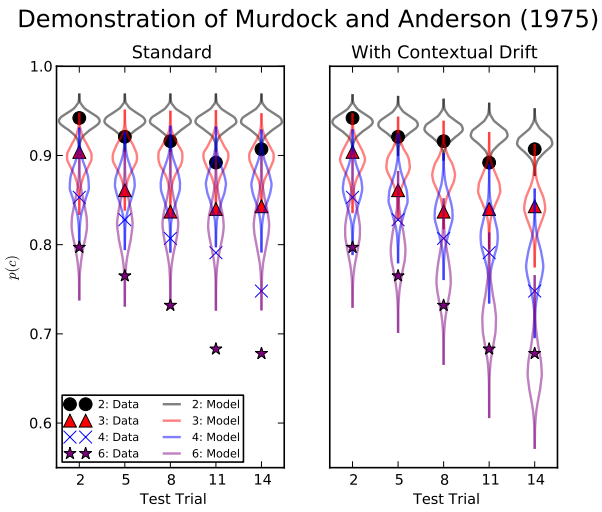

Figure 25. Group data and model predictions for the Murdock and Anderson (1975) paradigm, where the number of choices on a forced choice recognition test was manipulated between two and six. The left panel shows the model predictions without parameter modification, while the right panel shows the model's predictions each item on a test trial multiplies the mean match to context by .9985. Note: The depicted data reflect averages across serial positions on the study list.

Appendix A

Derivations for the Mean and Variance of Memory Strength in Associative Recognition

In associative recognition, the co-occurrence tensor M_c can be decomposed into the memories that are contributing to the memory strength s that is generated at retrieval from a list of pairs L. Pairs stored during the list episode are scaled by the associative learning rate r_{assoc}. Given that we are investigating datasets in which the word pairs are random pairings of unrelated words, we can assume that the probability of having seen a pair combination prior to the experiment is negligible and thus the context noise term can be omitted. We also assume no common words among the word pairs:

$$\begin{aligned} s_{intact} = (C'_s \otimes I'_a \otimes I'_b).r_{assoc}(C_s \otimes I_a \otimes I_b) + & \quad \text{Self Match} \\ \sum_{i,j \in L, i,j \neq a,b} r_{assoc}(C_s \otimes I_i \otimes I_j) + & \quad \text{Item Noise} \\ \sum_{u \in P, u \neq s} (C_u \otimes I_a \otimes I_b) & \quad \text{Background Noise} \end{aligned}$$

If the probe cue is a rearranged pair, there is no self match. However, there are two partial matches terms. For a list of pairs A-B, C-D, E-F, etc., a rearranged pair A-D will have a partial match to the stored A-B and C-D pairs:

$$\begin{aligned} s_{rearranged} = (C'_s \otimes I'_a \otimes I'_d).r_{assoc}(C_s \otimes I_a \otimes I_b) & \quad \text{Partial Match} \\ + r_{assoc}(C_s \otimes I_c \otimes I_d) & \quad \text{Partial Match} \\ + \sum_{i,j \in L, i,j \neq a,b,c,d} r_{assoc}(C_s \otimes I_i \otimes I_j) & \quad \text{Item Noise} \\ + \sum_{u \in P, u \neq s} (C_u \otimes I_y \otimes I_z) & \quad \text{Background Noise} \end{aligned}$$

The above equations can be rewritten as matches and mismatches between the item and context vectors:

$$
\begin{aligned}
s_{intact} = & r_{assoc}(C'_s.C_s)(I'_a.I_a)(I'_b.I_b)+ && \text{Self Match} \\
& \sum_{i,j \in L, i,j \neq a,b} r_{assoc}(C'_s.C_s)(I'_a.I_i)(I'_b.I_j)+ && \text{Item Noise} \\
& \sum_{u \in P, u \neq s} (C'_s.C_u)(I'_a.I_y)(I'_a.I_z) && \text{Background Noise}
\end{aligned} \tag{18}
$$

$$
\begin{aligned}
s_{rearranged} = & r_{assoc}(C'_s.C_s)(I'_a.I_a)(I'_d.I_b)+ && \text{Partial Match} \\
& r_{assoc}(C'_s.C_s)(I'_a.I_c)(I'_d.I_d)+ && \text{Partial Match} \\
& \sum_{i,j \in L, i,j \neq a,b,c,d} r_{assoc}(C'_s.C_s)(I'_a.I_i)(I'_b.I_j)+ && \text{Item Noise} \\
& \sum_{u \in P, u \neq s} (C'_s.C_u)(I'_a.I_y)(I'_a.I_z) && \text{Background Noise}
\end{aligned} \tag{19}
$$

The matches and mismatches of the context vectors can be substituted using the same parameters for the normal distribution used in Equation 7 for item recognition.

Appendix B

Equations for the Log Likelihood Ratio Transformation of Memory Strength

For our model, we used the analytic expressions for the likelihood ratio transformation derived by Glanzer et al. (2009) for both the equal variance and unequal variance normal distributions. In our model, the ratio of standard deviations $\sigma_{new}/\sigma_{old}$ depends on a number of model parameters, and ratios can be one or greater although ratios below one are the most common. Following Glanzer et al., we use X to refer to samples on the memory strength axis and Λ to refer to log likelihood ratios.

The equations described in Glanzer et al. (2009) describe a fully informed likelihood ratio model where the actual memory strengths are equivalent to the strengths used in the likelihood ratio. As described in the text, the expected strengths in the likelihood ratio need not be equivalent to the actual strengths. In the mixed lists in our list strength datasets, we assume that despite the fact that tested items have different strengths, the same expected strengths in the likelihood ratio are used for each item. We describe the expected strengths as "subjective strengths." We describe the actual means and standard deviations of the memory strength distributions as μ and σ, while the subjective means and standard deviations of the memory strength distributions are denoted as ρ and τ.

Glanzer et al.'s equations were written expressed for the case where $\sigma_{new} = 1$ and differences in strength were expressed as different values of μ_{old}. For that reason, in their expressions d' and d equal μ_{old} in the unequal variance and equal variance models, respectively. Given that in our model both σ_{new} and σ_{old} vary, all parameters are normalized by τ_{new} before usage in any of the equations below.

For the equal variance log likelihood ratio transformation, we use the equations in Appendix A of Glanzer et al. (2009). This results in normal distributions with means and variances as follows:

$$E(\Lambda|Old) = d'^2/2$$
$$E(\Lambda|New) = d'\mu_{old} - d'^2/2$$
$$Var(\Lambda) = d'^2\sigma_{old}^2$$

where $d' = \rho_{old}/\tau_{new}$. μ_{old} and σ_{old}^2 are normalized by τ_{new} and τ_{new}^2, respectively.

For the unequal variance log likelihood ratio transformation, we use the equations in Appendix B of Glanzer et al. (2009). Like with the equal variance case, σ_{new} is assumed to be one. The parameter ς refers to the subjective ratio of standard deviations τ_{old}/τ_{new}. The resulting log likelihood ratio distributions are non-central chisquare distributions, where the x is:

$$\frac{\Phi + (d^2)/[2\varsigma^2(\varsigma^2 - 1)] - (d^2/2\varsigma^2) - log(\varsigma)}{\sigma_X^2(\varsigma^2 - 1)/2\varsigma^2}$$

where Φ is the criterion on the log likelihood ratio axis and $d = \rho_{old}/\tau_{new}$, σ_X^2 is the actual variance of the target or lure distribution normalized by τ_{new}^2. The noncentral chisquare distribution has one degree of freedom and non-centrality:

$$(\frac{-d/(\varsigma^2/2) - \mu_X}{\sigma_X})^2$$

where μ_X and σ_X are the actual means and standard deviations of the target or lure distribution normalized by τ_{new}. Predictions for old or new items can be derived from these equations by substituting values of the appropriate distribution for μ_X and σ_X.

One should note that it is also possible for our model to produce cases where $\sigma_{new} > \sigma_{old}$ (although in the model fits, it was quite rare). Hit and false alarm rate predictions are produced for this case by merely taking using the cumulative distribution function with the parameters above, while in the standard $\sigma_{old} > \sigma_{new}$ case, hit and false alarm rates are produced by subtracting the cumulative distribution function from one.

Appendix C

Binomial and Multinomial Rate Estimation

While some investigations estimate the relevant parameters from signal detection theory (d' and c) and derive the hit rate h and false alarm rate f directly from those estimates (Dennis et al., 2008; Pooley et al., 2011), we did not want the rates to have to conform to a particular signal detection model. Instead, rates h and f were estimated directly from the hit and false alarm counts:

$$H_{i,j,k} \sim Binomial(h_{i,j,k}, T_{j,k})$$
$$F_{i,j,k} \sim Binomial(h_{i,j,k}, L_{j,k})$$

Confidence counts were sampled from a multinomial distribution:

$$H_{c1,i}, ..., H_{c6,i} \sim Multinomial(h_{c1,i}, ..., f_{c6,i}, T)$$
$$F_{c1,i}, ..., F_{c6,i} \sim Multinomial(f_{c1,i}, ..., f_{c6,i}, L)$$

All rates for individual participants in the yes/no conditions were sampled from reparameterized beta distributions, which uses a mean parameter λ and variance parameter ν:

$$h_{i,j,k} \sim Beta(\lambda_{h,j,k}, \nu_{f,j,k})$$
$$f_{i,j,k} \sim Beta(\lambda_{h,j,k}, \nu_{f,j,k})$$

Separate λ and ν parameters were used for each hit and false alarm rate for each condition. The hyperparameters for the means and variances used nearly non-informative priors:

$$\lambda_{j,k} \sim Beta(.5, 2)$$

$$\nu_{j,k} \sim InverseGamma(.1, .1)$$

For the confidence counts in the DeCarlo (2007) dataset, multinomial rates for each confidence category were sampled from reparameterized beta distributions, but were subsequently normalized to sum to one. We designate the samples before normalization as hc and fc:

$$hc_{c1,i}, ..., hc_{c6,i} \sim Multinomial(\lambda_{h,c1,i}, ..., f_{h,c6,i}, T)$$

$$fc_{c1,i}, ..., fc_{c6,i} \sim Multinomial(\lambda_{f,c1,i}, ..., f_{f,c6,i}, L)$$

Rates h and f were obtained by normalization:

$$h_{c1,i}, ..., h_{c6,i} = hc_{c1,i}/k, ..., hc_{c6,i}/k_{h,i}$$

$$f_{c1,i}, ..., f_{c6,i} = fc_{c1,i}/k, ..., fc_{c6,i}/k_{f,i}$$

where k is

$$k_{h,i} = k_{h,c1,i} + k_{h,c2,i} + ... + k_{h,c6,i} k_{f,i} = k_{f,c1,i} + k_{fc2,i} + ... + k_{f,c6,i} \tag{20}$$

Each rate parameter was estimated with four chains of 5,000 samples each after a burn-in period of 1,500 samples using JAGS software.

Appendix D

Cross Validation

To assess the generalizability of the model, we performed a cross validation procedure. In cross validation, model generalizability is assessed by fitting a model's parameters to a sample of the complete data while withholding some portion of remaining data. Subsequently, the model's fit is assessed by comparing its predictions to the withheld data. For the present purposes, we have adopted a k fold cross validation procedure, which has been shown to outperform the leave-one-out cross validation (LOOCV) method (Arlot & Celisse, 2010). In the k folds procedure, the data is equally divided into k sections, or folds, and the model is independently fit to each fold. To perform this procedure, we randomized the trial order of each participant's data and divided it up into eight folds ($k = 8$) due to the fact that eight was the largest number of observations for the associative recognition datasets. An equal number of observations for each condition and trial type were in each fold. The model was fit using the same prior distributions and number of chains (32) as in the original model fit.

For each fold in the cross validation, predictions for the withheld data were generated from the participant parameters in each fold from a single randomly selected set of parameters from each participant's posterior distribution. Subsequently, the predictions for each set of withheld data were summed together to produce a complete set of predictions for the withheld data. As an example, if there were eight observations per condition, a randomly selected set of parameters was used to generate predictions for the single withheld observation in each condition. For each fold there would be a single withheld observation, and the predictions from each of the eight folds were summed to produce a complete set of predictions (eight observations) for the withheld data for each participant.

To compare the predictions of the withheld data against the model, we used the rate estimation procedure described in Appendix C to estimate the posterior distributions on the hit and false alarm rates. The model's predictions were compared against the original

data in the same manner was used in the body of the text. Namely, posterior distributions on the group means for the hit and false alarm rates of both the data and the model were compared against each other. Depictions of the ROC and zROC data and predictions from DeCarlo (2007) can be seen in Figures D1 and D2, fits to the Dennis et al. (2008) dataset can be seen in Figure D3, fits to the data from the experiments using non-linguistic stimuli can be seen in Figure D4, and fits to the associative recognition data can be seen in Figure D5. One can see from inspection of the figures that the fit is quite good, and any impairment in the fit appears to be quite minor. This may be because several of the constraints on the model are operating across multiple datasets.

To assess the consistency in the parameter estimates across each of the folds, we calculated the proportion of total interference in the same manner as depicted in Figure 23. Median proportions of total interference for each interference contribution in each fold can be seen in Table D1. One can see that there is qualitative consistency across the eight folds. Each fold is in agreement with the conclusions of the main fit, namely that item noise is not a dominant source of interference in any of the datasets. Where there is the weakest across-fold consistency is in the two datasets that use fractals as stimuli. However, inspection of the interference proportion estimates of the main fit in Figure 23 reveals that the datasets that employ fractals as stimuli have the widest confidence intervals of all of the fits.

Table D1

Median proportions of total interference for each interference contribution (IN = item noise, CN = context noise, BN = background noise), as measured by each of the eight folds in the cross validation procedure.

	Folds							
Interference	1	2	3	4	5	6	7	8
Dennis et al. (2008): LF Words								

Continued on next page

Table D1 - continued from previous page

	Folds							
Interference	1	2	3	4	5	6	7	8
IN	.0008	.0027	.0004	.0007	.0041	.0004	.0019	.0029
CN	.0003	.0005	.0006	.0001	.0016	.0000	.0002	.0004
BN	.9955	.9888	.9936	.9961	.9841	.9971	.9917	.9872
Dennis et al. (2008): HF Words								
IN	.0004	.0014	.0002	.0004	.0021	.0002	.0009	.0015
CN	.0502	.4914	.4966	.4870	.4961	.4921	.4855	.4885
BN	.4928	.9888	.4977	.5089	.4959	.5048	.5064	.5012
List Length: Fractals								
IN	.1064	.2002	.1769	.0784	.2815	.2044	.1157	.1693
BN	.8934	.7998	.8231	.9215	.7184	.7955	.8842	.8306
List Strength: Fractals								
IN	.1583	.2888	.2475	.1207	.4141	.3132	.1724	.2441
BN	.8417	.7119	.7525	.8793	.5859	.6868	.8276	.7558
List Length: Faces								
IN	.3065	.3322	.3178	.3195	.2996	.2705	.3203	.2545
BN	.6935	.6678	.6821	.6804	.7000	.7294	.6797	.7455
List Strength: Faces								
IN	.3922	.4173	.3975	.4157	.3749	.3561	.3964	.3206
BN	.6088	.5827	.6025	.5842	.6250	.6438	.6035	.6793
List Length: Scenes								
IN	.0000	.0000	.0000	.0000	.0001	.0000	.0002	.0000
BN	.9999	.9999	.9999	.9999	.9999	.9999	.9998	.9999
List Strength: Scenes								

Continued on next page

Table D1 - continued from previous page

	Folds							
Interference	1	2	3	4	5	6	7	8
IN	.0000	.0000	.0000	.0000	.0002	.0000	.0003	.0000
BN	.9999	.9999	.9999	.9999	.9998	.9999	.9997	.9999
List Length: Pairs								
IN	.0000	.0000	.0000	.0000	.0001	.0000	.0000	.0000
BN	.9999	.9999	.9999	.9999	.9998	.9999	.9999	.9999
List Strength: Pairs								
IN	.0000	.0000	.0000	.0000	.0001	.0000	.0000	.0000
BN	.9999	.9999	.9999	.9999	.9998	.9999	.9999	.9999

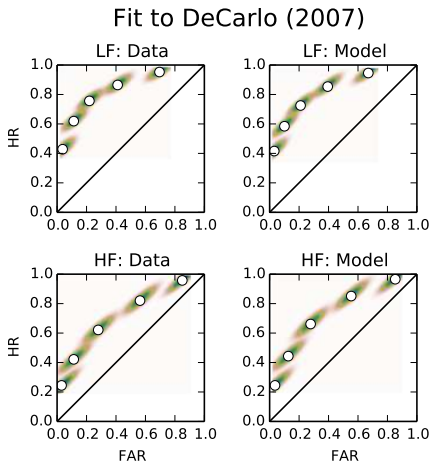

Figure D1. Density estimates of the ROC function for both the data of DeCarlo (2007, left) and cross validation model fit (right) for low (top) and high (bottom) frequency words. The circles indicate the median of the hit and false alarm rate posterior distributions for each ROC point.

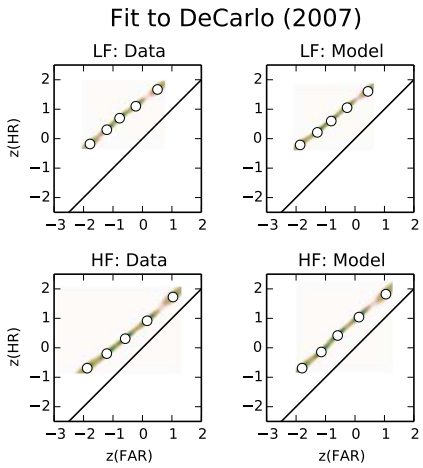

Figure D2. Density estimates of the zROC function for both the data of DeCarlo (2007, left) and cross validation model fit (right) for low (top) and high (bottom) frequency words. The circles indicate the median of the z-transformed hit and false alarm rate posterior distributions for each ROC point.

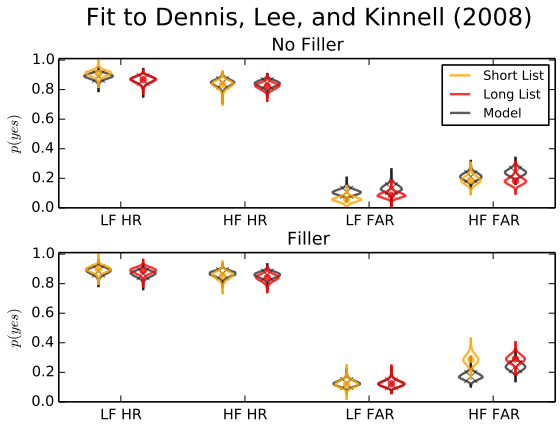

Figure D3. Fit to the data of Dennis et al. (2008) for all conditions. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's). Density estimates are depicted as teardrop plots, which vertically depict the entire posterior distribution by plotting them sideways (see the main text for more description).

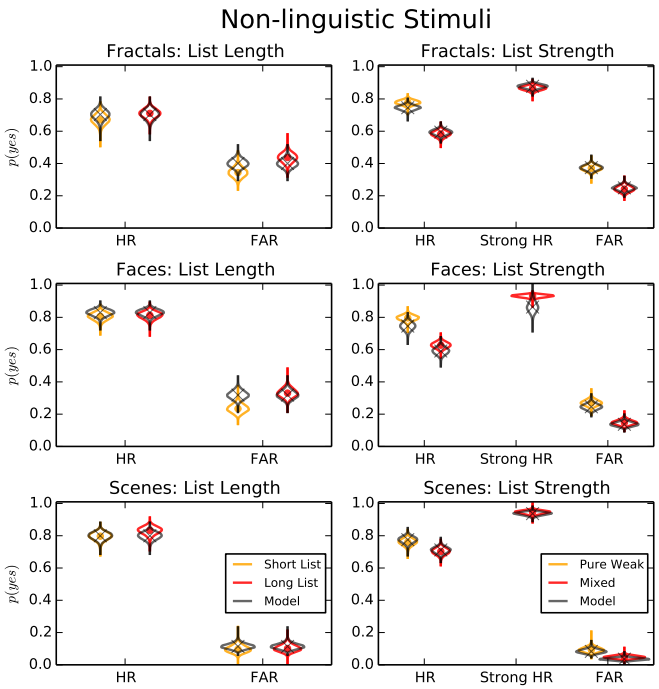

Figure D4. Hit and false alarm rates for the data and the cross validation model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth et al. (2014, right) that employ non-linguistic stimuli. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).

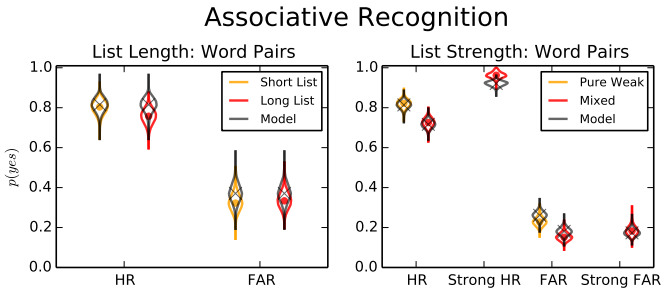

Figure D5. Hit and false alarm rates for the data and the cross validation model fit to the list length experiments of Kinnell and Dennis (2012, left) and the list strength experiments of Osth and Dennis (201, right) that employ word pairs in an associative recognition task. Depicted are the density estimates from group level parameters for both the data and the model, along with median posterior estimates for the data (circles) and the model (x's).