The list strength effect in source memory: Data and a global matching model

Adam F. Osth, Julian Fox, Meredith McKague

The University of Melbourne, Australia

Andrew Heathcote

The University of Tasmania, Australia

Simon Dennis

The University of Melbourne, Australia

Address correspondence to:

Adam Osth (E-mail: adamosth@gmail.com)

Author Note

This work was supported by a grant from the Australian Research Council, ARC DP150100272, awarded to Simon Dennis and Andrew Heathcote and ARC DE170100106 awarded to Adam Osth.

Abstract

A critical constraint on models of item recognition comes from the list strength paradigm, in which a proportion of items are strengthened to observe the effect on the non-strengthened items. In item recognition, it has been widely established that increasing list strength does not impair performance, in that performance of a set of items is unaffected by the strength of the other items on the list. However, to date the effects of list strength manipulations have not been measured in the source memory task. We conducted three source memory experiments where items studied in two sources were presented in a pure weak list, where all items were presented once, and a mixed list, where half of the items in both sources were presented four times. Each experiment varied the nature of the testing format. In Experiment 1, in which each study list was only tested on one task (item recognition or source memory), a list strength effect was found in source memory while a null effect was found for item recognition. Experiments 2 and 3 showed robust null list strength effects when either the test phase (Experiment 2) or the analysis (Experiment 3) was restricted to recognized items. An extension of the Osth and Dennis (2015) model was able to account for the results in both tasks in all experiments by assuming that unrecognized items elicit guess responses in the source memory task and that there was low interference among the studied items. The results were also found to be consistent with a variant of the retrieving effectively from memory model (REM; Shiffrin & Steyvers, 1997) that uses ensemble representations.

Keywords: recognition memory; source memory; global matching models

The list strength effect in source memory: Data and a global matching model

A major distinction in episodic memory research concerns the difference between information about learned content and the context in which it occurred. A common memory failure illustrating this distinction is that people remember a fact or detail but have no memory for where they learned the information. The relationship between memory for content and context is studied in the laboratory using the item recognition and source memory paradigms. In the item recognition paradigm, participants study a list of items and at test are asked to discriminate between studied items (targets) and unstudied items (lures). The source memory paradigm presents participants with a set of items in different sources, such as different font colors, studied locations, or sensory modalities. At test, participants judge which source studied items were presented in.

A number of computational models of decision making have been developed to explain the relations between item and source memory (e.g.; Banks, 2000; Batchelder & Riefer, 1990; DeCarlo, 2003; Yonelinas, 1999; Hautus, Macmillan, & Rotello, 2008; Glanzer, Hilford, & Kim, 2004; Klauer & Kellen, 2010; Slotnick & Dodson, 2005). These models fall into several frameworks including multivariate signal detection theory, in which participants make decisions based on continuous latent strengths (SDT: Banks, 2000), discrete state models (Batchelder & Riefer, 1990; Klauer & Kellen, 2010), or a combination of continuous latent strengths and discrete states (Yonelinas, 1999). The success of these models has generally been judged on their ability to account for the shapes of item and source memory receiver operating characteristics (ROCs). ROCs are usually constructed by instructing participants to make confidence ratings to items at test and calculating the cumulative hit rate (HR) and false alarm rate (FAR) across all confidence levels, beginning with the highest confidence response. The resulting ROC is typically curvilinear in item recognition (e.g. Egan, 1958), and is sometimes linear for source memory (Yonelinas, 1999), but becomes curvilinear when the ROC is constructed from only recognized items (Slotnick & Dodson, 2005). A number of models have been successful in accounting for this variety

of ROC shapes (Klauer & Kellen, 2010; Hautus et al., 2008; Yonelinas & Parks, 2007).

Murdock (2006) pointed out that a disadvantage of such models is that they make few predictions outside of ROC shapes. In particular, they are generally mute with respect to manipulations that often concern memory researchers, such as the effects of recency (Monsell, 1978), list length (Dennis, Lee, & Kinnell, 2008; Strong, 1912), list strength (Ratcliff, Clark, & Shiffrin, 1990), and word frequency (Glanzer & Adams, 1985), although, as addressed later, the Hautus et al. (2008) model makes one specific prediction with regard to the list strength paradigm in source memory. In contrast, the class of global matching models has generally been successful in explaining such effects (Clark & Gronlund, 1996). Unlike models in the SDT and discrete state frameworks, global matching models specify the processes underlying encoding and retrieval in episodic memory tasks, allowing them to provide more precise accounts of episodic memory phenomena. In global matching models, the similarity between the retrieval cues and each stored memory is computed. Subsequently, these similarities are summed together (or averaged; Shiffrin & Steyvers, 1997) to produce a single strength value that can be compared to a response criterion to make a decision. Collectively, the current crop of global matching models have been successful in explaining all of the aforementioned episodic memory phenomena in item recognition (e.g.; Dennis & Humphreys, 2001; Nosofsky, Little, Donkin, & Fific, 2011; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997), but have experienced little, if any, extension to source memory paradigms. The current article attempts to fill this gap by testing one of the major constraints on the development of global matching models, the list strength effect, in a source memory paradigm and introduces an extension of the Osth and Dennis (2015) model to account for the results.

The list strength paradigm: data and model predictions

A major prediction of early global matching models is that as the number of items in memory is increased, performance should decrease. In global matching models, each item

in memory has variation in its similarity to the retrieval cues, so that as the number of items in memory is increased, the number of variance components that contributes to the decision increases and the signal-to-noise ratio is reduced. Ratcliff et al. (1990) found that the models yielded the same predictions for the case of repetitions of the list items. This is because repetitions are treated in the same mannner as increases in the number of studied items; additional representations are added to memory, each of which increases noise at retrieval.

The models thus predicted that increasing the strength of a subset of studied items should impair performance on the remaining items. For instance, if one study list consisted of once presented items, while another contained a mixed-strength composition where half the items were presneted once (1x) and the other half of the items were presented four times (4x), the models predict that performance for once presented items should be worse in the mixed list of 1x and 4x items due to the extra interference caused by the repetitions. The list with more repeated items would be considered a list with higher "list strength."

A number of experiments tested and disconfirmed this prediction: increasing the strength of a set of studied items does not impair performance of the other items on the list for the case of item recognition with word stimuli (Kahana, Rizzuto, & Schneider, 2005; Ratcliff et al., 1990; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Shiffrin, Huber, & Marinelli, 1995; Yonelinas, Hockley, & Murdock, 1992) although small effects of list strength have been found with non-word stimuli such as faces and fractals (Norman, Tepe, Nyhus, & Curran, 2008; Osth, Dennis, & Kinnell, 2014). One should note that the free recall task contrasts with recognition memory in that increasing list strength has been shown to substantially impair performance (Malmberg & Shiffrin, 2005; Tulving & Hastie, 1972). Consequently, amendments to the global matching framework were proposed that enabled the models to predict a null list strength effect in item recognition. One such modification was the differentiation hypothesis, in which repetitions accumulate into a single strong memory trace that is more responsive to its own cue but less responsive

to other cues (Shiffrin, Ratcliff, & Clark, 1990). The latter component implies that strong memory traces generate less interference, whereas in the older models the opposite was the case.

Another class of models has argued that the null list strength effect is more indicative of a general lack of interference among studied items (Dennis & Humphreys, 2001; Murdock & Kahana, 1993a, 1993b; Osth & Dennis, 2015). While most models have assumed that memory is a blank slate before presentation of the study list, these models instead assume there is a large interference contribution from pre-experimentally learned memories. This interference can come from prior occurrences of the cue word (context noise) or from other memories in general (background noise). When such interference contributions are present, interference from the additional repetitions in a list strength paradigm produces only a negligible increase in overall interference, allowing the models to predict null effects of list strength.

To our knowledge, none of these models which have been successful in addressing benchmark phenomena in item recognition have been applied to the source memory task. Perhaps the simplest extension of these models to source memory would involve a binding between each item and its source at study; at test the probe item would be cued with each of the studied sources and the memory strengths of each source cue would be compared. As an example, consider if a participant studied a set of items such as "truck" and "joker" in source A and "sky" and "phone" in source B. At test, when prompted with a cue such as "truck", in order to make a judgment as to which source "truck" was studied in participants could cue memory with a binding of "truck" in source A and match it to the contents of memory to obtain the memory strength for source A. Subsequently (or in parallel), the participant could cue memory with a binding of "truck" in source B and match it to the contents of memory to obtain a memory strength for source B. The difference between the memory strengths for source A and B could be used to make a decision; the source that elicits the greater degree of memory strength would be chosen. Heathcote, Raymond, and

Dunn (2006) suggested a similar mechanism operates in the list discrimination paradigm, in which participants are asked to indicate which of two lists a test item was studied in, and the plurality discrimination paradigm, in which participants discriminate between studied items and switched-plurality lures (e.g., if participants studied "cat" and "kings", switched plurality lures would be "cats" and "king").

Although this mechanism is similar to item recognition, the representational structure of the memory set in the source memory and related tasks can lead to different empirical and theoretical results. In item recognition, a word such as "truck" receives its strongest contribution from its own representation in memory, while the other items on the list produce much smaller degrees of match, due to the fact that they bear little resemblance to the retrieval cue. However, in the source memory case, half of the items in the list match the source cue, meaning that source memory can resemble cases where half the studied list items bear high similarity to a retrieval cue. This is also the case in the plurality discrimination paradigm, where it has been found that list strength impairs discrimination (Norman, 2002; Buratto & Lamberts, 2008).

On the theoretical front we found that the higher similarity in the source memory paradigm was sufficient to induce a list strength effect in the original version of the retrieving effectively from memory model (REM: Shiffrin & Steyvers, 1997). This was somewhat surprising because in REM strengthening items produces differentiation of the memory traces, which should reduce the interference contribution from strong memory traces. Differentiation only reduces interference when the similarity between the trace and the cue is relatively low. In cases where this similarity is high, interference increases with strength (Criss, 2006). In source memory, half of the items have high similarity to the cue by virtue of the matching source features. When some of these items are strengthened, they will produce extra interference due to their strong match to the retrieval cues on the source dimension, and a list strength effect can be predicted. However, later formulations of REM allow additional ensemble features that are unique to a binding between items or

features (Criss & Shiffrin, 2005). We found that, when when sufficiently common, these ensemble features mitigate the interfering effect of the matching source features and a null list strength effect is predicted.

In the Osth and Dennis (2015) model, when the item-source cue is matched to memory, the similarity to each item-source binding in memory is a multiplication of the similarity to the item cue and a similarity to the source cue. Larger similarities to item-source bindings that mismatch the retrieval cues will produce larger interference and reduce performance. Previous investigations with the model have suggested that the matches to the other items on the list are low in magnitude (Osth & Dennis, 2015). We found that under such conditions, the model similarly predicts a null list strength effect in source memory. These predictions, along with the model, are detailed after we describe our empirical investigation and its results.

The Current Investigation

To our knowledge, the effects of list strength on source memory performance have not been investigated. A number of investigations have focused on the effects of strengthening a single source on the slope (e.g.; repeating source A items but not source B items) of the z-transformed ROC (Starns & Ksander, 2016; Starns, Pazzaglia, Rotello, Hautus, & Macmillan, 2013; Yonelinas & Parks, 2007). However, these studies did not explore the extent to which the strength manipulation impaired memory for the items that were not strengthened, which is the focus of the list strength design.

In each of our experiments, participants studied a list of 32 items, where half the items were presented in the lower left corner of the screen in a colored font (source A) and the other half of the items were studied in the upper right corner in a different colored font (source B). In the *pure weak* condition, each item was presented once. In the *mixed* condition, half the items were presented once (1x) and the other half were presented four times (4x). An equal number of source A and source B items were strengthened, and each

repetition was always presented in the same source. Participants were tested on item recognition and source memory in each experiment. We additionally manipulated word frequency in our experiments to place extra constraint on the computational model. Several prior investigations have found source memory advantages for low frequency words (Glanzer et al., 2004; Guttentag & Carroll, 1994, 1997; Marsh, Cook, & Hicks, 2006; Mulligan & Osborn, 2009), just as in item recognition.

We hypothesized item recognition should show no list strength effect, in that d' in the pure weak list should be equivalent to d' for weak items in the mixed list, because this null list strength effect in item recognition has been replicated quite extensively in the literature. Predictions for the source memory list strength effect are less clear. On one hand, an investigation conducted by Glanzer et al. (2004) found that a large number of conventional manipulations, such as word frequency and depth of processing, have similar effects on both item recognition and source memory, which might suggest that source memory should similarly show no effect of list strength. On the other hand, list strength effects have been found in some recognition paradigms that resemble source memory, such as in the plurality discrimination paradigm (Buratto & Lamberts, 2008; Norman, 2002). In addition, the list strength predictions from the REM model depend somewhat on the representational assumptions of the model, with the basic model predicting a list strength effect in source memory.

Although participants were tested on item recognition and source memory for both conditions in each of the experiments, the specific details of the testing varied somewhat from experiment to experiment. In Experiment 1, participants were tested on either item recognition or source memory for each studied list, but not both. Specifically, after completing the study list, they were post-cued on the test type. This experiment found a list strength effect in source memory (poorer source d' for 1x items in the mixed relative to the pure weak list) but not in item recognition. Although this result might seem to support the basic REM model, which predicts that the source features should provide extra

interference, Hautus et al. (2008) provide an alternative explanation in terms of decision processes. In their model, source memory judgments are only elicited for recognised items, while guesses are elicited by unrecognised items. Consistent with several other studies (e.g., Hirshman, 1995; Stretch & Wixted, 1998; Criss, 2006; Osth & Dennis, 2014; Osth et al., 2014; Starns, White, & Ratcliff, 2010), in Experiment 1 the item recognition HR was lower for 1x items in the mixed than in weak list [1]. Under the Hautus et al. (2008) model, the lower recognition rate for the mixed list would cause to more source memory guesses for 1x items than in the pure weak list, leading to poorer performance in the mixed list for reasons that are unrelated to interference from memory retrieval.

The Hautus et al. (2008) model predicts that if the source memory test was restricted only to items that were recognized by the participants, the list strength effect in source memory should be reduced or eliminated. This prediction was tested in Experiment 2, which was nearly identical to Experiment 1 with the exception that it used a conditionalized source memory procedure. During the test phase of Experiment 2, for each item, participants were initially tested on their item recognition; if they gave a "yes" response to an item, they were then immediately prompted for a source memory judgment. Items that were not recognized did not receive source memory judgments. No list strength effect was observed, which suggests that the list strength effect observed in Experiment 1 was due to the source guessing for unrecognized items.

A third experiment followed up this suggestions, using testing of item recognition and source memory for each studied item, but in separate phases. Participants were given an item recognition test after the study list and then were subsequently given a source memory test on all the studied items in a separate block. In addition, while Experiments 1 and 2 used two choice tests ("yes" vs. "no" for item recognition, "source A" vs. "source B"

[1]The reason why null list strength effects were observed in terms of d́despite decreases in HR is an accompanying decrease in FAR, a result which has been interpreted as due to a criterion shift (Hirshman, 1995; Stretch & Wixted, 1998)

for source memory tests), Experiment 3 used six point confidence ratings. This procedure enabled a post hoc conditionalization of the source memory data based on the confidence in the recognition responses. Again, no list strength effect was observed when source memory was restricted to recognized items.

Following description of the three experiments and their theoretical implications, we present the source memory extension of the Osth and Dennis (2015) model, and describe its application to all three experiments using hierarchical Bayesian techniques, which enables fitting of the individual participants while allowing for group-level constraints across each of the experiments.

Experiment 1

In Experiment 1, we tested both item recognition and source memory for the presence of a list strength effect. Words were presented in one of two sources. There were two source dimensions, color (green or yellow) and screen location (the lower left corner or the top right corner). The two source dimensions were correlated for each participant to improve source discriminability. Participants were presented with either a pure weak list, which contained all once presented items, or a mixed list, in which half the items were presented once and half the items were presented four times. Of the once and four times presented items, half were in one source and the other half in the other source. Following each study list, participants were tested on either item recognition or source memory for that study list.

Participants

Participants were 81 first-year psychology students at the University of Melbourne who received course credit.

Materials

A set of high ($N = 252$, CELEX frequency 100-560 occurrences per million) and low ($N = 341$, CELEX frequency 1-2 occurrences per million) frequency words were used for this experiment. These sets were drawn from the MRC Psycholinguistic Database and ranged from 5-9 letters and 1-2 syllables in length. All plurals or derivational variants of words were excluded. Both sets of words were approximately equated for the mean number of neighbors using N-Watch (Davis, 2005). The number of neighbors ranged from 0-12 for the high frequency words ($M = 1.988$, $SD = 2.395$) and from 0-8 for the low frequency words ($M = .968$, $SD = 1.425$).

Procedure

A diagram of the basic procedure can be seen in Figure 1. On each iteration, participants underwent either the pure weak condition or the mixed list condition. During the study phase of both conditions, participants studied a set of 32 words, where half were high frequency (HF) and half were low frequency (LF). Each word was presented on the screen for 2000 ms. Presentation of each word was followed by a blank screen for 250 ms. To engage their attention, participants were asked to indicate whether the word was pleasant or not using the "i" and "k" keys, respectively. Half the words were present in one source (source A) and half were in another (source B). Each source was composed of two source dimensions: color (green or yellow) and screen location (bottom left or upper right corner). For each participant, a color was randomly assigned to one of the screen locations. That is, either participants were presented with green words in the lower left corner and yellow words in the upper right corner, or yellow words in the lower left corner and green words in the upper right corner. Usage of two correlated source dimensions was intended to increase source discriminability, as we found it was quite poor when only screen location was manipulated for the two sources.

In the pure weak condition, each word was presented once. In the mixed list

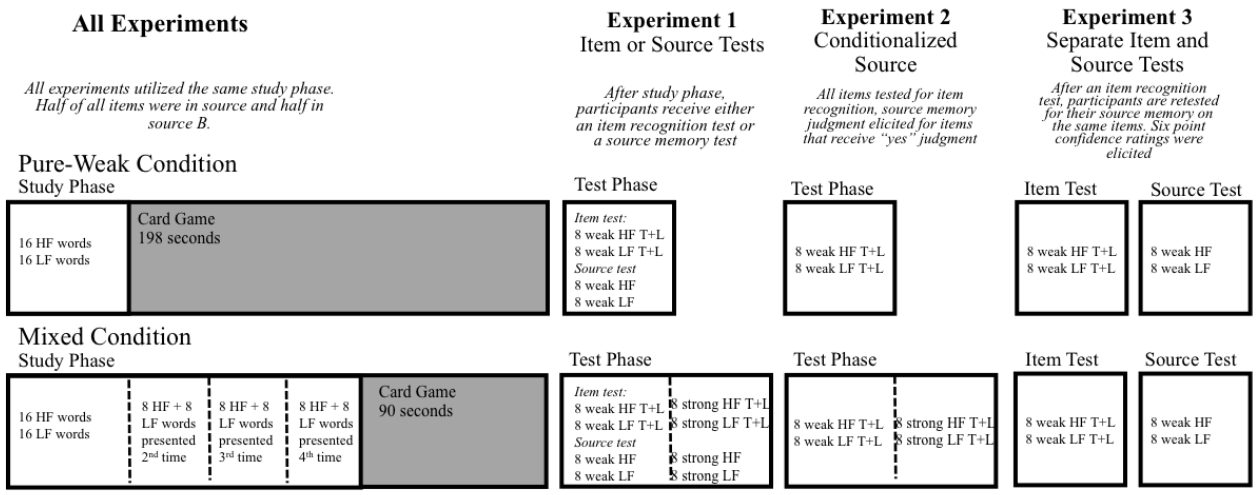

Figure 1. Diagram of the experimental procedure and how the test phase differs across each experiment. Notes: HF = high frequency, LF = low frequency, T = targets, L = lures.

condition, after all words were presented once, half of the words were presented three more times. The repetitions were blocked, in that each of the strengthened words had to be presented twice before they were presented on their third time, etc. An advantage of this design is that participants see all of the words on the mixed list before they know the words are repeated, which prevents them from differentially rehearsing the strong items at the expense of the weak items (rehearsal borrowing, e.g., Ratcliff et al., 1990). Of the repeated words on the mixed list, there were an equal number of HF and LF words and an equal number of words in source A and source B.

After the study phase, participants underwent a demanding distracter task in which they played a game where playing cards appeared on the screen at a rapid pace and they had to periodically make responses according to a set of rules, such as pressing the space

bar when two cards with the same suit appeared in a row. To encourage participation in the task, participants scored points for correct key presses and lost points for incorrect presses. The card game lasted for 90 seconds in the pure weak condition and 198 seconds for the mixed list condition. The purpose of the different lengths was to ensure that the retention intervals for weak words were identical in both conditions.

During the test phase for a given list, participants were tested on either item recognition or source memory for a given study list. Participants were post-cued with which task they would perform; they were not told until the instructions prior to the test phase. In all item recognition tests, participants were presented with test lists of which half of the items were targets and half were lures. For both targets and lures, half of the items were LF and half HF words and for targets, half the items were studied in source A and half in source B. Participants were instructed that they were to press the "1" key to indicate that they recognized studied items and "0" to indicate that they did not recognize the item; response buttons appeared on the screen throughout the test phase to remind them of the response keys. Source memory test lists were similar with the exception that there were no lures present on the study lists, making the test lists half as long as the item recognition test lists. During the source memory tests, participants were instructed to respond "1" to items in the lower left corner and "0" to items in the upper right corner. Response buttons were present throughout the test list to remind the participants of the response keys; these buttons were in the same color as the studied sources to additionally remind participants of the color dimension in their source judgments. The buttons were also arranged in a similar position as their studied locations; the source A button was presented on the lower left of the bottom half of the screen while the source B button was on the upper right of the bottom half of the screen.

For the test lists of the pure weak condition, participants were tested on half the studied items. In the mixed list condition, participants were tested on all of the once presented items before they were tested on the strong items. This was to ensure that the

test position for once presented items was the same across both conditions, as performance in item recognition has been found to decline monotonically with increasing test position (Criss, Malmberg, & Shiffrin, 2011; Peixotto, 1947; Ratcliff & Murdock, 1976). This also ensures that the strength composition of the weak item test blocks was the same across both the pure weak and mixed list conditions, as the strength composition of a test list has been found to additionally influence the effects of test position (Kiliç, Criss, Malmberg, & Shiffrin, 2017). Pure weak test lists were composed of 32 for item recognition and 16 items in length for source memory tests, while mixed list test lists were composed of 64 trials for item recognition and 32 trials for source memory.

Participants completed a total of eight study-test cycles, half with item recognition testing and half with source memory testing. Half of the study-test cycles were in the pure weak condition while half were in the mixed list condition. For each task, participants completed an equal number of pure weak and mixed list conditions.

Results

Results from Experiment 1, along with the other two experiments, can be found in Figure 2. A weakness of the null hypothesis testing framework (NHST) is its inability to provide support for the null hypothesis (Wagenmakers, 2007). For this reason, we employed Bayesian ANOVAs and t-tests with JASP software to calculate the Bayes Factor, which indicates the change supported by the data of the relative evidence for the alternative hypothesis against the null hypothesis. A $BF_{10} > 1$ indicates increased evidence for the alternative hypothesis, a $BF_{10} < 1$ indicates greater evidence for the null hypothesis, while $BF_{10} = 1$ indicates no change. Assuming both hypotheses are equally likely before observing the data, by convention, a BF_{10} that is in the 1-3 range or the .33-1 range provides only "anecdotal" evidence for or against the null hypothesis, respectively, while BF_{10} in the 3-10 or .1-.33 range provides substantial evidence, and $BF_{10} > 10$ or $BF_{10} < .1$ provide strong evidence for or against the null hypothesis (Jeffreys, 1961).

For source memory, a hit was defined as a source A response to an item studied in source A, while a false alarm was defined as a source A response to an item studied in source B. HRs and FARs in source memory are depicted in Figure 2. Source memory analyses were restricted to d', while for item recognition HR and FAR were analyzed in addition to d' to analyze bias effects in response to list strength. To avoid infinite values of d', hit and false alarm rates were transformed by adding .5 to the hit and false alarm counts and 1 to the number of targets and lures before calculating d' (Snodgrass & Corwin, 1988). This was done only for the calculation of d'; all analyses on HR and FAR throughout the paper are on the raw, untransformed rates.

The item recognition results largely replicated the list strength findings that have been previously reported in the literature. There was no list strength effect, in that d' did not differ between the pure weak ($M = 2.06$) and mixed ($M = 2.03$), $BF_{10} = .136$, lists. Nonetheless, in accordance with previous results (e.g.; Hirshman, 1995), the HR and FAR were affected by the list strength manipulation. HR for once presented items were lower in the mixed list ($M = .791$) relative to the pure weak list ($M = .831$), $BF_{10} = 95.03$. There were also lower FAR in the mixed list ($M = .125$) relative to the pure weak list ($M = .164$), $BF_{10} = 8.30$.

We observed a word frequency effect; LF words exhibited much higher discriminability than HF words ($M_{LF} = 2.32, M_{HF} = 1.77$), $BF_{10} = 6.43e + 13$. The locus of the LF advantage was primarily in the FAR, where there were large differences between LF ($M = .091$) and HF ($M = .199$) words, $BF_{10} = 1.94e + 15$, while the HR advantage for LF words was not as strong ($M_{LF} = .833$, $M_{HF} = .795$), $BF_{10} = 11.41$. Item recognition HRs for words presented in each of the two sources was nearly identical ($M_{left} = .804$, $M_{right} = .805$), $BF_{10} = .092$, which provides suggestive evidence that the two sources were represented with equal strength in memory.

In the source memory task, a robust list strength effect was observed, as reflected in poorer discriminability in the mixed list ($M = .945$) relative to the pure weak list

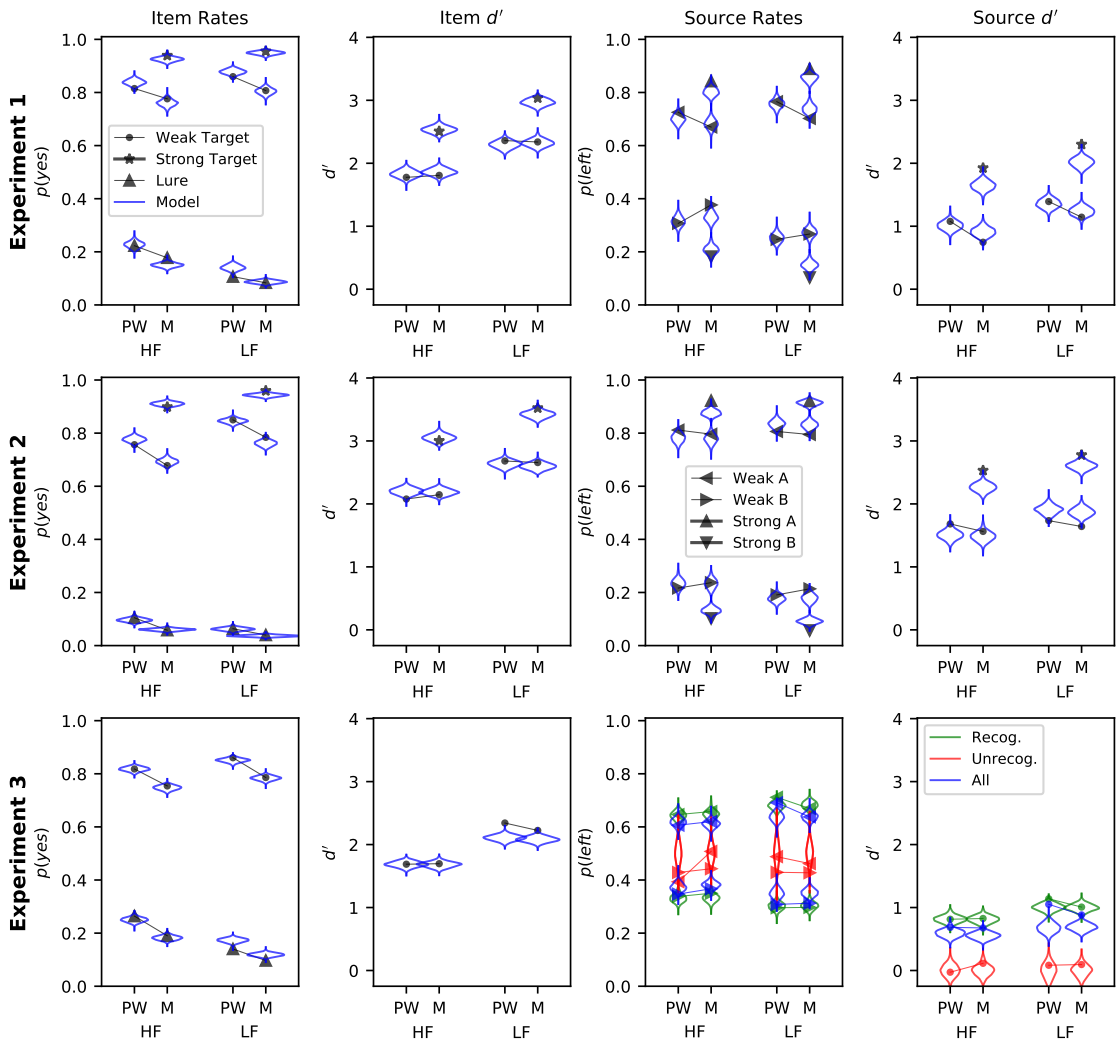

Figure 2. Group averages along with model predictions for Experiments 1 (top row), 2 (middle row), and 3 (bottom row) for the item recognition task (left two columns) and source memory tasks (right two columns). The posterior predictive distribution of the model is depicted using violin plots. Experiment 3 shows data and model predictions in the source memory task depending on whether the items were recognized (blue) or not recognized (red) in the item recognition task, in addition to showing all of the data/predictions (blue).

($M = 1.23$), $BF_{10} = 213.45$. Post hoc tests revealed that this is especially evident for HF words ($M_{PW} = 1.07$, $M_M = .745$), $BF_{10} = 152.46$. While a similar trend was evident for

LF words ($M_{PW} = 1.39$, $M_M = 1.14$), the Bayes Factor revealed only very weak evidence in favor of the effect, $BF_{10} = 1.14$. In addition to the list strength effect, there was a strong discriminability advantage for LF words ($M = 1.27$) over HF words ($M = .91$), $BF_{10} = 12,440$. The LF advantage also persisted for items that were presented four times ($M_{LF} = 2.30$, $M_{HF} = 1.93$), $BF_{10} = 25.28$

Discussion

Experiment 1 found a robust list strength effect on source memory performance under conditions where a null list strength effect in item recognition has been found. At first glance, these results support the predictions of the basic REM model, of a null list strength effect in item recognition but a list strength effect in source memory. This occurs as the representational structure in source memory is somewhat different from that of item recognition, due to half of the items bearing a strong resemblance to the source retrieval cue. In the basic version of the REM model, this implies that these items should exhibit higher interference with stronger learning (Criss, 2006).

Experiment 2 tested an alternative explanation from the two-dimensional signal detection model of Hautus et al. (2008). Although this model has been primarily applied to the shapes of the item recognition and source memory ROCs, it specifies a source memory decision mechanism that can produce list strength effects. Although source memory decisions are based on the degree of memory strength for a given source, an assessment of source strength is only made if an item is recognized; source memory judgments for unrecognized items instead elicit guess responses. This mechanism was motivated by analyses of the source memory ROC conditioned on different levels of item confidence. Performance generally increased as item recognition confidence was higher, but when items were unrecognized the source memory ROCs were often on the main diagonal (Slotnick & Dodson, 2005), indicating chance performance. Other investigations have similarly found source memory to be at chance for unrecognized items (Bell, Mieth, & Buchner, 2017;

Malejka & Broder, 2016).

In the item recognition data for Experiment 1, hit rates for once presented items were significantly lower in the mixed list relative to the pure weak list. The greater number of unrecognized items in the mixed list should imply that there are greater degrees of guessing on the source memory judgment, producing poorer source memory performance in conditions of higher list strength. Thus, the Hautus et al. (2008) model can predict a list strength effect in source memory without any decrease in source memory discriminability in the mixed list. If the decision criterion for item recognition is higher in the mixed list relative to the pure weak list, more guessing should occur on the mixed list. The data from Experiment 1 are relatively consistent with this idea in that the decrease in HR in the mixed list (.052) is quite close to the decrease in the percentage of correct source memory responses (.045). We would also like to explicitly note that this mechanism is not inconsistent with the aforementioned global matching models. Indeed, later in the article we implement it in the Osth and Dennis (2015) model.

Experiment 2

The explanation by the Hautus et al. (2008) model of the results of Experiment 1 can be tested as it predicts that, if source memory performance is conditionalized on item recognition, there should be no effect of list strength on source memory discriminability. Experiment 2, using an approach developed by Yonelinas (1999), affords such a test, as it is similar to the design of Experiment 1 with the exception that, for each item, participants had to give two types of decisions, first an old/new item recognition judgment, and then immediately after for items for which this was positive, a source A/source B judgment.

Participants

Participants were 78 first-year psychology students at the University of Melbourne that participated in exchange for course credit.

Materials

Materials were identical to Experiment 1.

Procedure

The procedure was identical to Experiment 1, with the exception that during each test list, participants were tested on both item recognition and source memory. For each item, participants were initially tested on item recognition in the same manner as in Experiment 1. If participants made an "old" response to any of the items, they were subsequently prompted for a source memory judgment in the same manner as Experiment 1. Due to the increased time required to complete the test lists, we reduced the total number of study-test cycles to six (three per list strength condition) from eight in Experiment 1.

Results

Results can be found in the middle row of Figure 2. The item recognition results are largely consistent with the results of Experiment 1 and prior literature. There was no list strength effect on d' ($M_{PW} = 2.38$, $M_M = 2.34$), $BF_{10} = .15$. HRs were lower in the mixed list than the pure weak list ($M_{PW} = .803$, $M_M = .731$), $BF_{10} = 55,087$, and FARs were lower in the mixed list as well ($M_{PW} = .084$, $M_M = .047$), $BF_{10} = 382,181$. LF words were more discriminable than HF words, $BF_{10} = 3.42e + 18$, as LF words exhibited a higher HR than HF words, $BF_{10} = 4.67e + 10$, along with lower FAR, $BF_{10} = 8559$. Item recognition HRs were again found to be virtually equal between the two sources, $BF_{10} = .120$.

The source memory results contrasted markedly with those from Experiment 1. There was no effect of list strength on discriminability ($M_{PW} = 1.79$, $M_M = 1.68$), $BF_{10} = .36$. Post hoc tests revealed that there was no effect of list strength on either HF words ($M_{PW} = 1.68$, $M_M = 1.56$), $BF_{10} = .144$, or LF words ($M_{PW} = 1.73$, $M_M = 1.64$), $BF_{10} = .173$. In addition, there was no effect of word frequency on discriminability in

source memory for once presented words ($M_{LF} = 1.77$, $M_{HF} = 1.73$), $BF_{10} = .183$. However, for strong words, there was an LF advantage ($M_{LF} = 2.77$, $M_{HF} = 2.53$, $BF_{10} = 9.37$.

Discussion

We conducted Experiment 2 as a test of a mechanism from the Hautus et al. (2008) model, in which guessing is elicited during source memory judgments when items are not recognized. This mechanism predicts poorer performance in conditions of higher list strength because the lower recognition rates of once presented items should produce more guessing on source memory decisions, and thus poorer performance. Therefore, in Experiment 2 we conditionalized source memory performance on item recognition by only allowing participants to make source memory judgments when items were recognized. Consistent with the Hautus et al. model, we found no effect of list strength on source memory discriminability.

We were somewhat surprised to find no effect of word frequency on source memory performance for once presented words. We initially hypothesized that the conditionalized procedure may have eliminated the effect, as it's possible that the effect in source memory stems from an item memory advantage. However, Glanzer et al. (2004) found an LF advantage in source memory using a similar conditionalized procedure. Nonetheless, there was an LF advantage for strong items. Given these inconsistencies, we will refrain from making strong conclusions about the absence of the word frequency effect for once presented words in this dataset.

Experiment 3

To further assess the generality of the results from Experiment 2, we conducted an additional experiment where each item receives both an item recognition judgment and a source memory judgment. However, in contrast to Experiment 2 where the two judgments were made in immediate sucession, participants made the judgments in separate blocks.

Specifically, participants engaged in an item recognition test list before beginning a source memory test list. In addition, six point confidence ratings were collected, which allows the conditionalization of source memory data on high confidence recognition responses in the item recognition task (Slotnick & Dodson, 2005).

Participants

Participants were 112 volunteers who were paid $10 for their participation in the study. They were recruited using online advertisements and printed flyers.

Materials

Materials were identical to Experiment 1 and 2.

Procedure

The procedure was similar to Experiments 1 and 2. Upon completion of the study phase, participants completed an item recognition test on 16 once presented targets and 16 lures. Unlike Experiment 1 and 2, the strong items were not tested in the mixed list condition. This is because the item recognition test list was followed by the source memory test, where they were tested for their source memory of the 16 once presented targets in a randomized order. If participants were tested on the strong targets in the item recognition test, it would increase the retention interval for the source memory test in the mixed list and potentially reduce performance.

Unlike Experiments 1 and 2, participants gave responses using a 6 point confidence scale. In the item recognition test, they were presented with buttons on the screen to remind them of the confidence keys, which included "1 = SURE OLD", "2 = PROB. OLD", "3 = UNSURE OLD", "8 = UNSURE NEW", "9 = PROB. NEW", "0 = SURE NEW". The source memory test used the same keys and confidence labels, but instead referred to the left and right sources, and the edges of the boxes corresponding to each source were colored in the same source as the studied sources.

Results

Our initial analysis of the results collapsed across confidence ratings; a response to a target was considered a hit, and a response to a lure was considered a false alarm, if the response was an "old" response, regardless of the confidence in the decision; the same assumptions were applied to the source memory responses. Three participants were excluded from all analyses and modeling for having extremely poor performance on the item recognition task ($d'_{item} < .15$); two of these participants had similarly poor performance on the source memory task.

The results can be found in the bottom row of Figure 2. The item recognition results replicate those of Experiment 2 and prior results in the literature. When discriminability was calculated as d', there was no list strength effect ($M_{PW} = 2.01$, $M_M = 1.96$), $BF_{10} = .156$. Both HR ($M_{PW} = .839$, $M_M = .771$), $BF_{10} = .87e + 10$, and FAR ($M_{PW} = .165$, $M_M = .126$), $BF_{10} = 21757$, was reduced in the mixed condition relative to the pure weak condition. Performance was better for LF than HF words, $BF_{10} = 2.71e + 24$. LF words exhibited higher HR, $BF_{10} = 160.96$, and lower FAR, $BF_{10} = 1.822e + 19$, than HF words. Item recognition HRs were virtually equivalent between the two sources, $BF_{10} = .075$.

For the source memory data, there was weak evidence for a null list strength effect ($M_{PW} = .868$, $M_M = .779$), $BF_{10} = .403$. Post hoc comparisons using Bayesian t tests revealed that there was no list strength effect for HF words ($M_{pw} = .685$, $M_m = .676$), $BF_{10} = .107$, while LF words showed weak evidence for a list strength effect ($M_{pw} = 1.05$, $M_m = .88$), $BF_{10} = 1.97$. There was a substantial word frequency effect, with LF words exhibiting higher d' ($M_{LF} = .967$, $M_{HF} = .681$), $BF_{10} = 328,799$.

We subsequently conditioned the source memory data by restricting it to items that were recognized during the item recognition test (targets that received an "unsure old" response or a higher level of confidence), which excluded 23.7% of source memory respones. These results are depicted in blue in Figure 2, while the source memory data for

unrecognized items are depicted in red. This restriction produced stronger evidence for a null list strength effect ($M_{pw} = .977$, $M_m = .912$), as evidenced by a lower BF, $BF_{10} = .171$. Post hoc tests revealed that while HF words still exhibited a null list strength effect ($M_{pw} = .814$, $M_m = .826$), $BF_{10} = .107$, LF words exhibited weak evidence for a null list strength effect ($M_{pw} = 1.14$, $M_m = 1.01$), $BF_{10} = .423$. A word frequency effect persisted after the conditionalization as well, although the evidence for the effect is much weaker than when the analysis is unrestricted ($M_{LF} = 1.14$, $M_{HF} = .820$), $BF_{10} = 798.84$.

We then restricted the source memory data to items that received a high confidence ("sure old") response during the item recognition task, which excluded 31.2% of the source memory responses. This restriction produced stronger evidence for a null list strength effect, $BF_{10} = .107$, and extremely similar d' scores in the pure weak ($M = 1.04$) and mixed list ($M = 1.03$) conditions. Both HF ($M_{pw} = .892$, $M_m = .951$), $BF_{10} = .132$, and LF ($M_{pw} = 1.18$, $M_m = 1.12$), $BF_{10} = .130$, words exhibited strong evidence for a null list strength effect. The word frequency effect again persisted in this analysis, although evidence for the effect is again weaker than in the previous analysis ($M_{LF} = 1.15$, $M_{HF} = .922$), $BF_{10} = 182.96$.

Source Memory for Unrecognized Items. Unrecognized items (depicted in red in Figure 2) showed source memory performance that was extremely close to chance. Due to a high proportion of recognized items (76.3%), several participants had insufficient data to calculate d' for each condition. In addition, participants varied considerably in their proportion of recognized items, and participants with very few unrecognized items produced extremely noisy estimates of d'. To partially ameliorate this problem, we collapsed across conditions while calculating d'. This analysis found that d' was extremely close to chance ($M = .15$) and the Bayes Factor produced only ambiguous evidence of being above chance ($BF_{10} = 1.13$).

There is a strong possibility that the varied number of observations per participant were responsible for the agnostic results. We found that participants with higher

proportions of unrecognized items produced values of d' that were much closer to zero, while participants with very low numbers of unrecognized items produced d' values that could be as extreme as 2 or -2 due to the small numbers of observations. In our Bayesian t-test, each of these observations are given the same weight despite the fact that there is much more uncertainty for participants with lower numbers of unrecognized items.

For this reason, we additionally analyzed our data using hierarchical Bayesian SDT models applied to the unrecognized items. For comparison purposes, we also ran the same models on the recognized items. Hierarchical Bayesian models are advantageous because they allow for the simultaneous estimation of group and participant level parameters. This allows for better estimation of the participant-level parameters, as they are constrained by the group level distribution, a phenomenon referred to as "shrinkage," which effectively reduces outliers and weights each persons estimate by its uncertianty. Additionally, a hierarchical Bayesian model naturally deals with missing observations by relying on group level information when data are missing. While space precludes a thorough treatment of hierarchical Bayesian models, interested readers should consult Lee (2011) and Rouder and Lu (2005).

In all models, we allowed a criterion for each confidence response and allowed criteria to vary across the list strength conditions, which accounted for ten parameters in each model. The models varied with their assumptions about the d' parameter. In the simplest model, d' was fixed to zero (the $d' = 0$ model); only decision criteria were estimated for this model. The subsequent models allowed for varying degrees of factoring of the d' parameter, including a single d' across all conditions, d' varying over word frequency conditions (d' ~WF), d' varying over list strength conditions (d' ~LS), and d' varying over all conditions (d' ~LS, WF). Posterior sampling was accomplished using differential evolution Markov chain Monte Carlo (DE-MCMC) sampling, a technique which is robust to correlations among parameters (Turner, Sederberg, Brown, & Steyvers, 2013). For each model, the number of chains was set equal to three times the number of parameters. Sampling began

after 6,000 burn-in iterations were discarded. The chains were thinned such that only one in every 20 samples was collected; this process continued until 2,000 samples were collected in each chain. Relatively non-informative prior distributions were employed for the model parameters; these are described in Appendix B.

Each model was compared using the widely applicable information criterion (WAIC: Watanabe, 2010), a metric which imposes a complexity penalty. In WAIC, model complexity is measured by the variability in the likelihood of a data point across posterior samples summed across all data points and is an approximation to leave-out-one cross validation. Smaller values of WAIC mean that a model gives better out-of-sample predictions by striking a balance between goodness-of-fit and simplicity. Because WAIC is on a log likelihood scale, differences between models by 10 points are conventionally considered large. We additionally calculated the conditional probability of each model using the weighting recommended by Wagenmakers and Farrell (2004). These results can be found in Table 1

Table 1

WAIC values and conditional probabilities for each hierarchical SDT model applied to the unrecognized and recognized items from Experiment 3. N = number of parameters per participant.

		Unrecognized		Recognized	
Model	N	WAIC	Prob.	WAIC	Prob.
$d' = 0$	10	4375	0	11827	0
Single d'	11	4365	.003	9714	0
d' ~LS	12	4372	0	9795	0
d' ~WF	**12**	**4353**	**.9997**	**9653**	**1.0**
d' ~LS, WF	14	4370	.0005	9759	0

The results strongly reject the $d' = 0$ model, indicating source memory for unrecognized items. The preferred model for both unrecognized and recognized items shows that d' only varies across HF and LF words but does not vary across the list strength conditions. The posterior distribution for the group means of d'_{HF} and d'_{LF} can be seen in Figure 3. d' is above chance for both HF (M = .098, 95% highest density interval, or HDI: [.019, .17]) and LF (M = .208, 95% HDI: [.089, .325]) words. These estimates stand in stark contrast to the d' estimates for recognized items, which are considerably higher for both HF (M = .82, 95% HDI: [.70, .95]) and LF (M = 1.16, 95% HDI: [1.03, 1.31]) words. Thus, while source memory appears to be above chance for unrecognized items, it is nonetheless extremely close to chance, and is much poorer than source memory for recognized items.

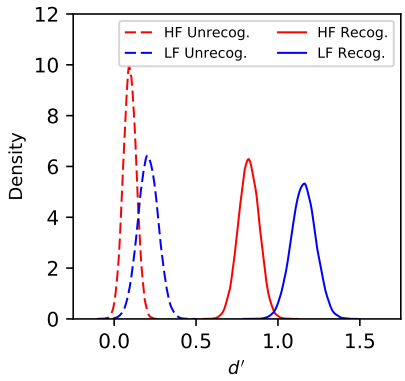

Figure 3. Posterior distributions of the group means of the d' parameters from the d' ~WF SDT model applied to the unrecognized items from Experiment 3.

Discussion

When the source memory data were analyzed across all responses, regardless of whether the items were recognized, the results showed only weak evidence for a null list strength effect. This stood in contrast to the results of Experiment 1, which found significantly poorer performance in the mixed list condition relative to the pure weak

condition. Although there was a significant decline in the HR as list strength increased (.069), the decrease in the HR was bigger with increasing list strength was bigger than the decrease in the percentage of correct source memory responses (.016). Source guessing on unrecognized items would predict an equivalent decrease in both tasks. However, when the data were conditioned on recognized items and again conditioned on high confidence item responses, stronger evidence for a null list strength effect in source memory was obtained. This was especially the case for source memory trials where the items were recognized with high confidence, which produced virtually identical d's for the pure weak and mixed list condition. This occurred because source memory for unrecognized items was extremely close to chance performance and there were more unrecognized items in the mixed list. Thus, when these items were removed from the source memory analysis, it equated the performance across the pure weak and mixed list conditions to a greater degree than when then unrecognized items were included in the analysis.

When we applied a hierarchical Bayesian SDT model to our data, we observed slightly above chance source memory performance for unrecognized items, which is in somewhat of a contradiction to the prior literature (e.g.; Malejka & Broder, 2016). Starns, Hicks, Brown, and Martin (2008) found above chance source memory for unrecognized items, but only under conditions of very conservative responding (when participants were told that only 25% of tested items were new). Nonetheless, source memory performance on unrecognized items was still extremely close to chance performance. In addition, several of the previous analyses relied on null hypothesis significance testing (NHST) and did not apply hierarchical Bayesian models to the data, which address the different numbers of responses from the participants in a principled fashion by appropriately weighting individual estimates according to variations in uncertainty when calculating population estimates.

Although the above-chance performance for unrecognized items may appear challenging for the mechanism in the Hautus et al. (2008) model, which states that source guessing occurs on unrecognized items, there are two ways that the model may be

reconciled with the current data. First, is that the presentations of the items in the item recognition test facilitated their recognition on the source memory test, making them slightly less likely to elicit guess responses on the source memory test. This explanation seems somewhat unlikely given that source memory performance was considerably poorer in this experiment compared to the other two experiments. The other possibility is that there is criterion variability in the old-new decision (e.g.; Benjamin, Diaz, & Wee, 2009; Osth, Bora, Dennis, & Heathcote, 2017), meaning that when participants assess whether an item is old on the source test, it's possible that they do so with a different decision criterion than they employed on the initial item recognition test. This means that some items that elicited "new" responses on the initial item recognition test elicited an implicit "yes" response on the source memory test, which prompted source retrieval instead of a guess. In the next section, we evaluate the feasibility of the source guessing mechanism of the Hautus et al. (2008) model within a computational model of item recognition and source memory.

A Global Matching Model of Source Memory

Here, we introduce a new computational model of source memory to explain the present data. The model is an extension of the Osth and Dennis (2015) model to source memory. Unlike models in the framework of SDT or discrete states, our model describes the representations that underlie the task and specifies the retrieval process. Just as with item recognition, source memory is described as a global matching process, whereby the cues on the test trial are matched against all of the contents of memory, producing a summed memory strength that reflects the similarity of the cues to the contents of memory (Clark & Gronlund, 1996).

Two factors distinguish source memory from item recognition. First, source memory is cued with each source cue (source A and source B) and the difference between each source cue's memory strength is compared to a decision criterion to produce a decision. Second, the item cue only matches one representation from the study list, whereas the

source cue matches half the items on the study list, producing additional interference. The reason why additional interference is produced is that the degree of interference is proportional to the strength of the match in global matching models (Osth & Dennis, 2015; Shiffrin et al., 1990).

The Osth and Dennis (2015) item recognition model storedmassociations between items and contexts. The term "context" in episodic memory models is fairly broad, but tends to refer to a representation that defines the episode. In episodic recognition tasks, participants are not asked if they've ever seen the items on the test list; if they were, the answer to any familiar item would be "yes." Instead, participants are asked whether they've seen the item in a particular episode, namely the study list. Thus, in a majority of episodic memory models, there is a context representation that defines the study list episode; the extent to which this context representation is different from prior contexts determines the extent to which participants are protected from interference from episodes prior to the study list (Dennis & Humphreys, 2001; Klein, Shiffrin, & Criss, 2007; Lohnas, Polyn, & Kahana, 2015). Although episodic memory models are often agnostic as to what defines context, a candidate explanation is the participant's cognitive and/or emotional state along with aspects of their physical surroundings.

In our extension of the model to source memory, we describe source A and source B as source contexts that are separate from the episodic context corresponding to the study list. The reason the source context is separate from the episodic context is that the source manipulations often used in source memory tasks, such as different font colors, spatial locations on a computer screen, or modalities of presentation are insufficient by themselves to define an episode; from the source information alone, one would not be able to deduce whether or not an item was in the current list or a previously studied list. In our model, the items (I), episodic contexts (C), and source contexts (S) are each defined using separate vectors and are combined into a conjunctive representation. We formalize this conjunctive representation as a mode three tensor, which is a three-way outer product of

the I, C, and S vectors:

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

where r is a learning rate parameter. The subscript s indicates that the context vector corresponds to the study episode, subscript t indicates the item vector is an item from the list, subscript a denotes that the source context corresponds to source A, and the set L corresponds to the items on the study list. Memory strength (s) is determined by combining the context cue at retrieval, the item cue during the test list, and a representation of one of the source contexts:

$$s = (C'_s \otimes I'_t \otimes S'_a).M \tag{2}$$

where the dashes indicate that the cues employed may not be identical to the vectors stored at study. Conventional applications of the tensor model proceeded by generating vectors from sampling distributions with a finite number of elements. Our model circumvents this approach by using an approximate analytic solution that specifies the similarities between the vectors without specifying the content of the vectors. The derivations of the model and equations for the means and variances of the memory strength distributions for item recognition and source memory can be found in Appendix A.

The mathematics of the model are conceptually illustrated in Figure 4. The similary between a cue on a particular dimension (item, context, or source) and a component in memory is specified as a normal distribution. Each dimension's similarity is multiplied by the other dimension's similarities, resulting in a multiplication of the similarity of the item, context, and source dimensions. This is done for all memories and the similarities are subsequently summed together. All memories that are not the target item contribute additional variance to the memory strength distributions and reduce the signal-to-noise ratio. One should note that Figure 4 simplifies the model by only demonstrating the source A cue; subsequently, the source B cue is applied and the difference between the two

memory strengths is calculated.

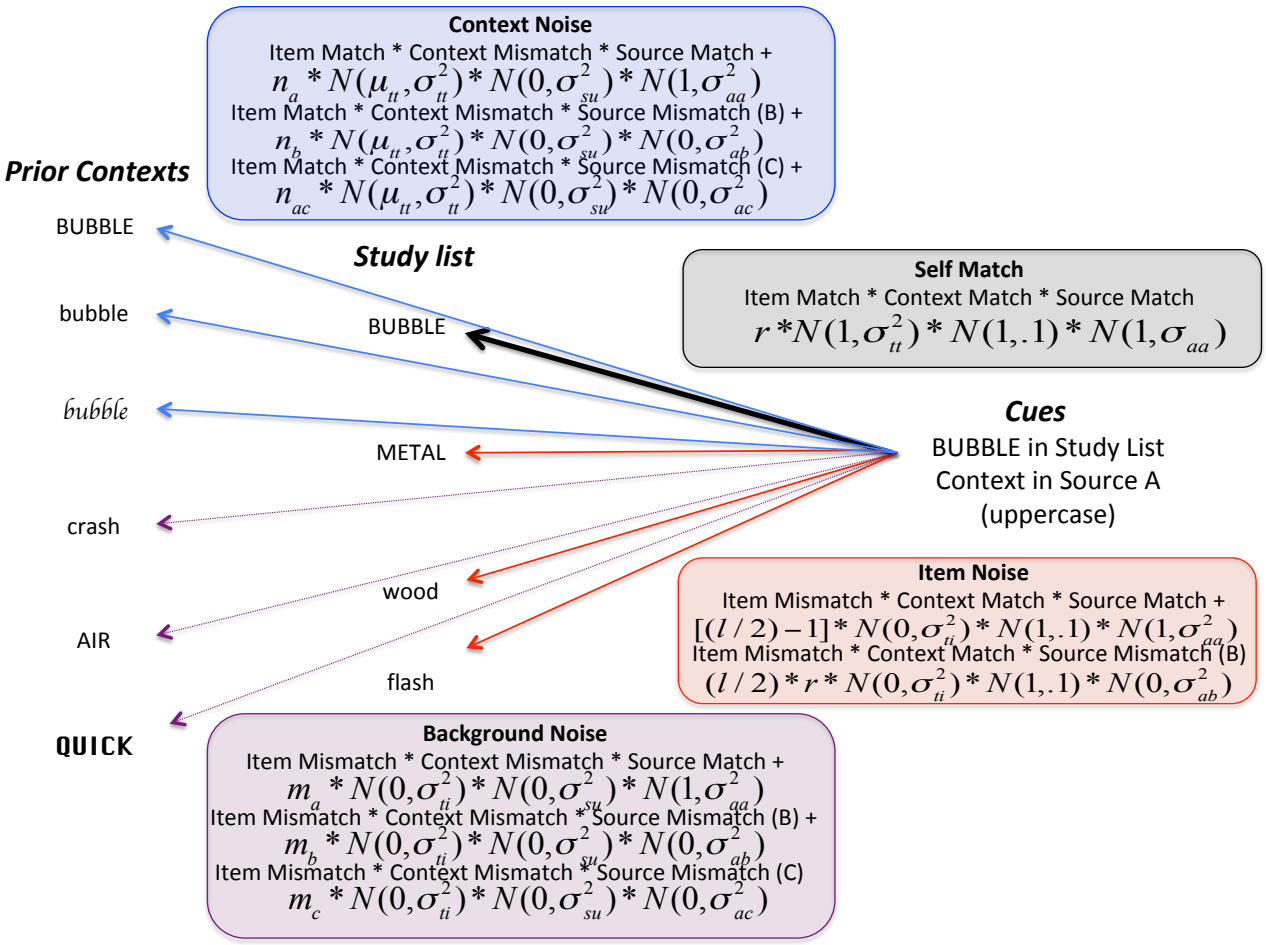

Figure 4. Illustrative example of the source memory extension of the Osth and Dennis (2015) model. Items in memory are associated with either the study list (right) or prior contexts (left). Items in memory were also studied in source A (denoted by uppercase font), source B (denoted by lowercase font), and other sources (denoted by other fonts). A cue comprising an item cue ("bubble"), a context cue, and the source A cue are globally matched against each item in memory. Each box represents a different interference category. See the main text for details and the Appendix for the mathematical implementation.

The items in memory fall into four categories depending on whether they match or

mismatch the item and context cues. Figure 4 illustrates an example where the word "bubble" is used as a cue, along with a context cue that represents the study list and the source A cue. In the example, source A was illustrated using uppercase letters, source B as lowercase letters, and other sources for memories acquired prior to the list episode are depicted using other less conventional fonts. Since "bubble" was a studied item, there is a binding between "bubble" and a representation of the study list context in source A present in memory. This is referred to as the *self match*, and is the primary determinant of performance in item recognition because lure cues do not match any of the items on the list. In source memory, both the source A and source B cues are used to probe memory. There is a self match present for each of these cues, but one of them will mismatch the source cue.

The match on the item dimension is a draw from a normal distribution with a mean equal to 1 and variance σ_{tt}^2, which is a parameter of the model. The match on the context dimension is a draw from a normal distribution with mean 1 and variance .1. The match on the context dimension has a psychological interpretation which corresponds to the participant's ability to reinstate the study list context or the extent to which the context has "drifted" from the context that had been initially learned (Osth & Dennis, 2015). However, in all of our experiments the retention interval was not manipulated, so it was necessary to fix these parameters to conventional values.

The mean of the self match is primarily determined by the learning rate r, which increases with study time and/or repetitions. The variability of the self match is determined by the item match variability parameter σ_{tt}^2. Psychologically, this parameter might correspond to variability in encoding an item's features from presentation to presentation (e.g., McClelland & Chappell, 1998). As σ_{tt}^2 is increased, the variability of the target distribution is increased relative to the lure distribution, which allows for the predictions of zROC slopes that are less than one in item recognition (Osth & Dennis, 2015). The match on the source dimension is a draw from a normal distribution with mean 1 and variance σ_{aa}^2, which is an additional parameter of the model that reflects the noise in

the match of a source to its own representation. As mentioned previously, if this parameter is sufficiently high, there will be a strong degree of interference from the items in memory that were bound to the source A context.

Item noise refers to the penalty from items that were present on the study list but mismatch the item cue, such as the word "wood" in Figure 4. The mismatch on the item cue is a sample from a normal distribution with mean 0 and variance σ_{ti}^2. Increases in the item mismatch variability parameter σ_{ti}^2 increase the noise contribution from other items on the list, and decrease performance as a result. Psychologically, this can correspond to the degree to which items are similar to each other; if items are completely dissimilar to each other, there will be no item mismatch penalty. The source mismatch is a sample from a normal distribution with mean 0 and variance σ_{ab}^2, which increases the noise contribution from sources that mismatch the source cue. The source memory case differentiates itself from the predictions for item recognition because in source memory, the magnitude of item noise depends on whether the stored source information matches or mismatches the source cue, with matching source cues producing more interference. Because each item that is not the target item contributes variance, the signal-to-noise ratio is reduced as the number of items is increased (where the number of study list items is denoted by l) or as the variance from each item mismatch (the σ_{ti}^2 parameter) is increased.

Item noise increases with increases in the learning rate r, which is what is responsible for the prediction of a list strength effect. While increases in the learning rate increase the contribution from the target item (increasing the signal-to-noise ratio), they also increase the noise penalty from other items on the list. To understand why that is, consider the fact that if r decreases toward zero for the other list items, it would eventually reach the point where they are no longer in memory and cannot produce interference. However, one should also note that this critically depends on the item mismatch penalty parameter σ_{ti}^2. As this parameter decreases toward zero, the penalty for mismatching other items on the list decreases toward zero for both item recognition and source memory. Osth and Dennis

(2015) demonstrated that the list strength effect of varying magnitudes could be produced with different values of σ_{ti}^2 for different stimuli, with words exhibiting very low item mismatch penalties and confusable stimuli such as fractal images exhibiting relatively high values.

Context noise refers to the penalty from memories that match the item cue but were learned prior to the experiment. A word such as "bubble" has been experienced by a participant many times over the course of their lifetime. "Bubble" was also likely to have been experienced in various source contexts - these source contexts might include source A and B along with many other sources that were not manipulated in the experiment, such as different sensory modalities, locations, or speakers; in Figure 4 these are depicted using non-conventional fonts. The penalty for mismatching the context representation is a draw from a normal distribution with mean 0 and variance σ_{su}^2 which is multiplied by the number of prior occurrences of the item. In item recognition, context noise predicts poorer performance on HF words, as their greater number of prior occurrences produces a larger memory strength penalty (e.g., Dennis & Humphreys, 2001). One should note that the same predictions can apply to source memory here - items that have been experienced in more sources prior to the experiment should exhibit poorer performance, which can produce advantages for LF words in source memory.

Background noise refers to the penalty from memories that were learned prior to the experiment that mismatch the item cue. Background noise comprises interference from all other unrelated memories acquired across the participant's lifetime and was previously incorporated into the TODAM model (Murdock & Kahana, 1993a, 1993b; Murdock, 1997).

The tensor model described above applies to source memory; how might predictions be derived for item recognition, given that the memory structure is a mode three tensor? We follow Humphreys, Bain, and Pike (1989) and assume that when undergoing item recognition, participants attempt to cue their memory without any reference to source information. We did this by assuming a generalized source cue, which has a match of one

to all source vectors and no variance (see Appendix A for more details). Because this cue exhibits noise in source memory, the interference contribution can be much larger in source memory than in item recognition.

A complete list of model parameters used in the model fit is depicted in Table 2. Several model parameters were fixed to improve parameter estimation and because they were not consequential to the performance of the model; these are described in Appendix B.

Table 2

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

Param	Bounds	Description
r_{weak}	$0 : \infty$	Learning rate for once presented items.
rs	$1 : \infty$	Strength factor for strong items; multiplied by r_{weak} to produce learning rate for strong items (r_{strong}).
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.
m_{HF}	$0 : \infty$	Number of prior occurrences of HF items in memory. Increases context noise for HF words.
σ^2_{ti}	$0 : \infty$	Item mismatch variability: Increases item and background noise in the model.
σ^2_{su}	$0 : \infty$	Context mismatch variability: Increases context and background noise in the model.
σ^2_{aa}	$0 : \infty$	Source match variability: Increases noise for matches on the source dimension. Note that $\sigma^2_{bb} = \sigma^2_{aa}$
σ^2_{ab}	$0 : \infty$	Source mismatch variability: Increases noise for mismatches on the source dimension. Note that $\sigma^2_{ba} = \sigma^2_{ab}$

Continued on next page

Table 1 - continued from previous page

Param	Bounds	Description
σ^2_{ac}	$0 : \infty$	Source mismatch variability: Increases noise for mismatches for sources outside of the experiment. Note that $\sigma^2_{ac} = \sigma^2_{bc}$
Φ_{item}	$-\infty : \infty$	Response criterion for item recognition. 0 represents an unbiased criterion.
Φ_{source}	$-\infty : \infty$	Response criterion for source memory. 0 represents an unbiased criterion.

Prior to the decision, the memory strength distributions for both item recognition and source memory are subjected to a log likelihood ratio transformation (Glanzer, Hilford, & Maloney, 2009) using the linear approximation developed by Osth, Dennis, and Heathcote (2017). The essence of a log likelihood ratio transformation is that memory strength is compared to an expected degree of memory strength for a given condition; conditions with higher expectations are held to a higher standard, which results in lower log likelihood ratios. This produces the mirror effect (Glanzer & Adams, 1985, 1990), because conditions of better performance are held to a higher standard, reducing the FAR in item recognition. This is critical for the list strength predictions here. In the mixed lists, the strong items are compared to higher retrieval expectations; specifically a degree of learning that is the average of the weak and strong learning rates. These higher expectations predict that HR and FAR should be reduced in the mixed list, just as is found in data. Analytics for the log likelihood ratio distributions of the model can be found at the end of Appendix A.

Predictions for the paradigm can be seen in Figure 5. Predictions in item recognition (HR and FAR in the left panel, d' in the middle panel) and source memory (d''; right panel) were generated with a range of different values of the item mismatch variability

parameter σ^2_{ti}, which governs the total amount of item noise, along with a range of different values of the source match variability parameter σ^2_{aa} and source mismatch variability parameter σ^2_{ab}, which increase the interference of the source representations. To simplify the predictions, a fixed background noise of .05 was assumed for item recognition and .1 for source memory. A greater degree of background noise was employed for source memory due to the fact that background memories will produce more interference from the source cues. All predictions were generated for a set of 16 items learned with a learning rate of 1.0. The other half of the study list items were not tested but were added to memory with a learning rate r_2 that was varied between .01 and 4.0, to evaluate the extent to which the learning rate of these items interfered with the other half of the list items.

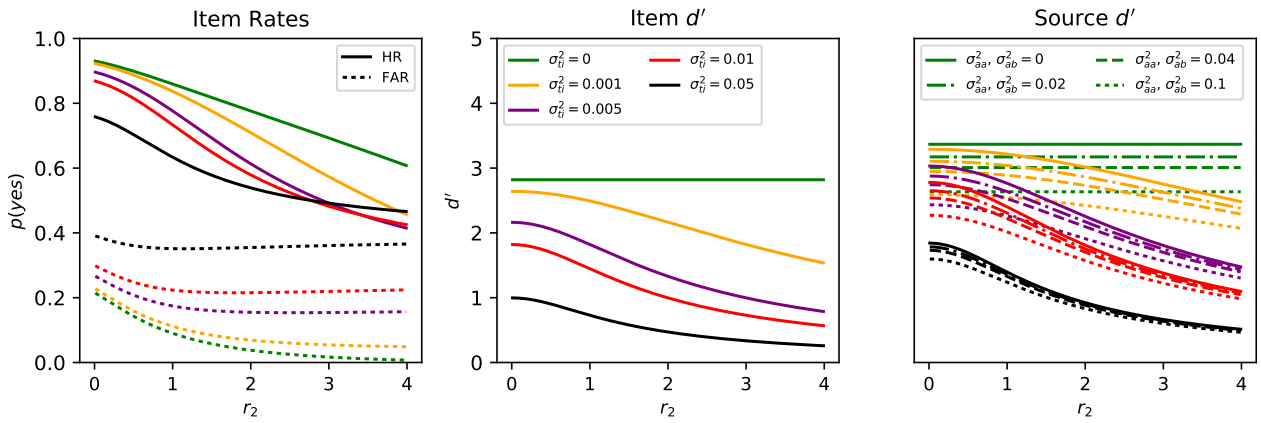

Figure 5. Model predictions for the list strength paradigm for item recognition (HR/FAR in left panel, d' in right panel), and source memory (d'; right panel). See the main text for more details. Model parameters: $r = 1.0$, $\sigma^2_{ss} = .1$, $\sigma^2_{tt} = .01$, $\sigma^2_{su} = .015$, $\sigma^2_{ac} = .25$, $n = 5$, $\Phi_{item} = 0$, $\Phi_{source} = 0$.

One can see from the figure that when $\sigma^2_{ti} = 0$ (green lines), which is the case where there is no item noise, d' is completely unaffected by the learning rate of the other items, meaning that no list strength effect is predicted in item recognition or source memory. This is evident in the middle and right panels, where the model's d' is unchanged by the

strength of the second set of items. This applies in source memory even when as the parameters that govern interference from the source representations (σ^2_{aa} and σ^2_{ab}) is increased (the dashed and dotted lines show highter values of σ^2_{aa} and σ^2_{ab}). It does not, however, mean that r_2 has no effect on performance, as the HR and FAR in item recognition (left panel) decrease as r_2 is increases. As σ^2_{ti} is increased above zero, a list strength effect in both item recognition and source memory is predicted as r_2 is also increased; this is evident as a decrease in d' as r_2 is increased for all models where $\sigma^2_{ti} > 0$. Interestingly, the parameters that govern interference among the sources (σ^2_{aa} and σ^2_{ab}) do not appear to strongly interact with the σ^2_{ti}, implying that the list strength predictions are mostly reliant on the σ^2_{ti} parameter and not on the parameters governing interference from the source representations.

The Model Fit

The model was fit to all three experiments simultaneously using hierarchical Bayesian analysis. Parameters that were allowed to vary across experiments were given separate group-level distributions, so that data from other experiments have no influence on those parameters. Wherever possible, however, we attempted to use a single group level distribution across all experiments to provide a strong degree of constraint on the model.

The only parameters that were allowed to vary across experiments were the item and source criteria along with the learning rate for once presented items, r_{weak}. The learning rate was varied across experiments to capture the differing degrees of performance in each. Experiment 2 had the best performance, while Experiment 3 showed substantially worse performance in source memory. This could be because the differing nature of the test formats in each experiment encouraged different degrees of learning from the participants. Alternatively, the poor performance in Experiment 3 could be due to the fact that the source memory test occurred after the item test, which would result in a weaker match to the study list context, or due to greater criterion variability as a consequence of the

six-point confidence ratings (e.g.; Benjamin, Tullis, & Lee, 2013). Distinguishing between these different possibilities in the model would add little for the present purposes.

Specific details of the hierarchical model implementation, such as the prior distributions on the model parameters, are described in Appendix B. The model employed 11 parameters per participant for Experiments 1 and 2, 19 parameters per participant for Experiment 3, and 25 pairs (mean and standard deviation) of group level parameters. Although this might seem like a lot of parameters, one should note that the only parameter to vary across the pure weak and mixed list conditions is the additional learning rate parameter for strong items. All other parameters are held constant across the two conditions. In addition, several parameters corresponding to the source cues (σ^2_{aa}, σ^2_{ab}, and σ^2_{ac}) could potentially be fixed in future applications of the model based on the parameter estimates here. All models were run with 75 chains; after 20,000 burn-in iterations were discarded the chains were heavily thinned such that one in every 20 samples was kept until 1,500 samples per chain were collected.

We fit two implementations of the model that varied with their assumptions about the data from Experiment 1. The first model assumes that source memory responses in Experiment 1 are a latent mixture of guesses and retrieval from source memory, a model we refer to as the *mixture* model. More specifically, the likelihood of source responses was calculated according to the model itself and according to a guessing process where all source memory decision probabilities were fixed to .5. The weight of the likelihood of the guesses was one minus the old-new hit rate for that condition while the weight of the likelihood of the informed decisions was weighted by the old-new hit rate for that condition. This model applies the mechanism of the (Hautus et al., 2008) model; the psychological interpretation of the latent mixture is that participants do not bother to attempt source retrieval when they do not recognize the items.

For Experiment 2, all of the responses were assumed to be informed source responses because participants only gave source memory responses for items they had recognized. For

Experiment 3, the recognized and unrecognized items are known due to participants having been tested on both. The recognized items were assumed to be source informed while the unrecognized items were assumed to be guesses. For the guessing, the probabilities of each confidence response were fixed to 1/6.

An additional model was fit that assumed that there was no latent mixture of guesses and source informed decisions, a model we refer to as the *non-mixture* model. This model instead assumed that all source memory responses came directly from the model; no guessing process was included. Both models had the same number of parameters; while mixture models often require additional parameters for mixing rates (e.g.; DeCarlo, 2002), in our mixture model the mixing parameter is determined by the old-new hit rate in item recognition, making this model variant quite constrained.

Posterior predictive distributions from both models were generated by simulating a dataset for 5% of posterior samples (one in every 20 posterior samples) from each participant and averaging across participants. Figure 6 shows the results from each model for Experiment 1. The posterior predictive distribution is depicted using violin plots, where greater width represents greater probability mass of the predictions in that region. Both models exhibit very similar predictions with the exception of the list strength effect. In particular, the non-mixture model notably fails to predict the list strength effect; equivalent performance is predicted for the pure weak and mixed list conditions. This is likely because the item recognition data along with the null list strength effect in the other two experiments offer strong constraint on the model parameters that prevent the model from predicting a list strength effect. The mixture model, however, predicts poorer performance in the mixed list relative to the pure weak list. This is because in the item recognition predictions, there is a lower hit rate to the weak items in the mixed list relative to the pure weak list, which produces a greater degree of guessing in the mixed list and poorer performance. Aside from the source memory data in Experiment 1 and the non-mixture model's inability to predict the differences between recognized and unrecognized items in

Experiment 3, the two models yielded extremely similar predictions. For this reason, all other predictions were generated using the mixture model; these can be seen in Figure 2.

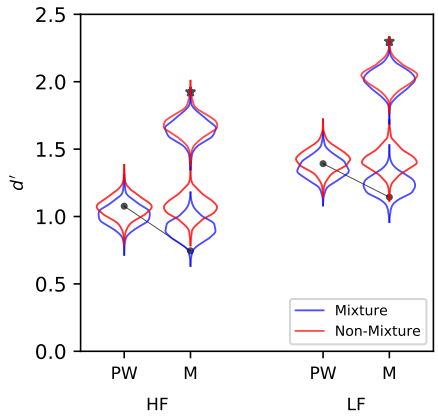

Figure 6. Group averages along with model predictions for the source memory d' in Experiment 1 for the mixture and non-mixture models.

We additionally compared the models on quantitative grounds by comparing the WAIC scores from each model. The mixture model (WAIC = 39,251) improved over the non-mixture model (WAIC = 39,301) by 50 points; a substantial improvement. Outside of Experiment 1, all model predictions are derived from the mixture model to provide cleaner figures. Nonetheless, outside of the source memory data in Experiment 1 the two models yielded nearly identical predictions.

Outside of the list strength effect in Experiment 1, the model does a very good job of addressing the data from each experiment. The model is also successful in predicting null effects of list strength in item recognition (all experiments) along with the null effect of list strength on source memory in Experiment 2. In Experiment 3, the model predicts a smaller effect of list strength when the predictions are restricted to recognized items, as that has the effect of removing the greater number of unrecognized items in the mixed list than the pure weak list. It is also interesting to note that while the hierarchical Bayesian SDT model identified that source memory performance was slightly above chance for unrecognized items in Experiment 3, the source memory d's for unrecognized items did not

fall outside our model's posterior predictive distribution (lower right panel in Figure 2). The model also appears to be providing a strong account of the word frequency effect in source memory. One notable exception is that in Experiment 2 the model predicts a LF advantage in source memory for once presented items, while the data showed no effect. However, it remains unclear why Experiment 2 shows no LF advantage for once presented items, as both Experiments 1 and 3 show an LF advantage, and in Experiment 3 the advantage remains when the analysis is restricted to recognized items. Another limitation is that the model appears to somewhat underpredict the performance on strong items in source memory for each experiment.

For item recognition the model gave a very good account of the data, and was able to account for all of the qualitative trends; this included the LF advantage in each experiment along with the reduced HR and FAR in the mixed list relative to the pure weak list. The latter predictions follow from the likelihood ratio decision mechanism (Glanzer et al., 2009; Osth, Dennis, & Heathcote, 2017), which produce a higher standard for evidence in conditions of higher list strength. Although we have demonstrated these phenomena with this model previously (Osth & Dennis, 2015), the work here demonstrates that the predictions persisted when the model was jointly constrained by the source memory task.

ROC predictions and group averaged data for Experiment 3 can be seen in Figure 7 for item recognition (top two panels), along with source memory for recognized items (second row), unrecognized items (third row), and collapsed across recognized and unrecognized items (fourth row). Model predictions are the mean of the posterior predictive distribution. For item recognition, the mixed list condition yields a similar shape as the pure weak condition but is shifted to the left due to the more conservative responding (reduced HR and FAR). For source memory, the ROC shapes for the pure weak and mixed list conditions are extremely similar to each other. One area of misfit is the fact that for unrecognized items, the data are slightly above the diagonal indicating slightly above chance performance, while the model is on the diagonal due to guessing. The

model's fit here could potentially be improved by implementing a mechanism whereby previously unrecognized items are recognized on the source test due to criterion variability or due to variability in the processing of the item. This would, however, add additional complexity to the model. Aside from this misfit, the model appears to be providing a good account of the ROCs across tasks and conditions.

To get a sense of why the null list strength effect was captured by our model, we calculated the magnitude of each interference category, namely the self match, item noise, context noise, and background noise, in item recognition and source memory. Because two source cues are employed in source memory, we separated the self match and item noise contributions depending on whether there was a match on the source dimension (same source, e.g.; item was studied in source A and A was used as a cue) or a mismatch on the source dimension (different source, e.g.; item was studied in source A and source B was used as a cue). The proportions of each interference contribution are depicted in Figure 8. We restricted consideration to the mixed list condition for each task because that condition exhibits the highest degree of item noise. In addition, we restricted consideration to Experiment 1 because the other experiments yielded nearly identical results.

One can see in the figure that in both tasks, item noise never dominates the interference contributions. In fact, its proportional contribution to the source memory task appears to be much smaller than the other sources, where background noise effectively dominates. Nonetheless, one can see that for both the self match and item noise in source memory, the interference is much larger for the items in memory that match the source cue rather than the ones that mismatch the source cue. The results of the computational modeling suggest that in both item recognition and source memory, null list strength effects are found because item noise, which increases with increasing list strength, plays a relatively small role in both item recognition and source memory.

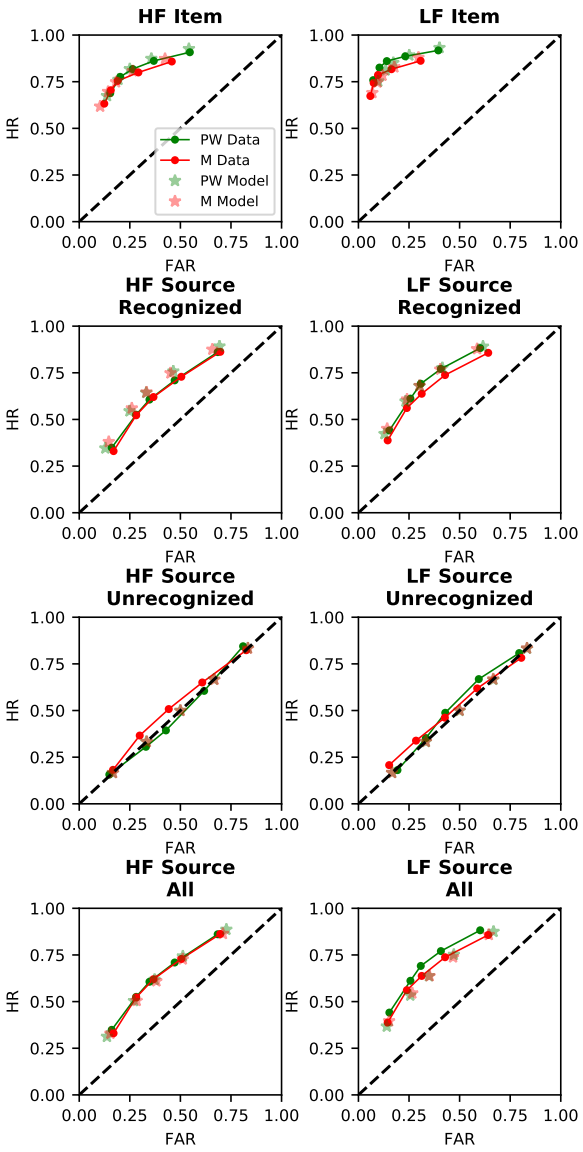

Figure 7. Group averaged ROC data and model predictions for Experiment 3 for item recognition (top row), along with source memory for recognized items (second row), unrecognized items (third row), and collapsed across recognized and unrecognized items.

General Discussion

In three experiments, we tested for the presence of a list strength effect in source memory. In Experiment 1, where participants were not tested on both item recognition and source memory on the same items, we found strong evidence for the presence of a list

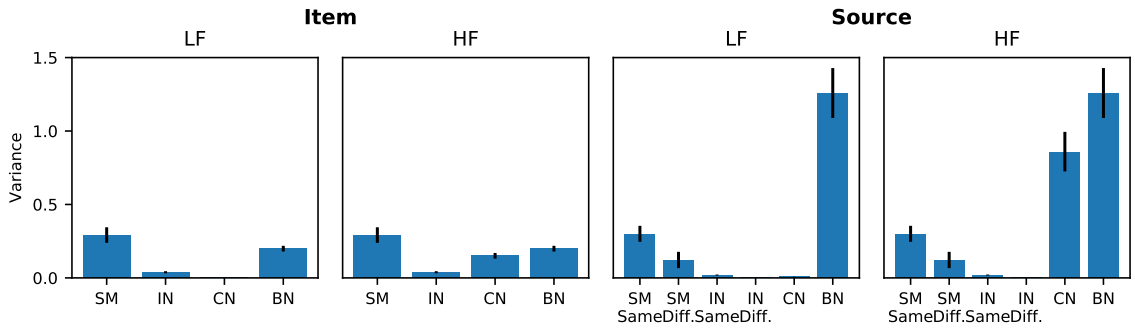

Figure 8. Memory strength variance from each interference category in item recognition (left two panels) and source memory (right two panels). Note that SM = self match, IN = item noise, CN = context noise, and BN = background noise. In source memory, the self match and item noise components are divided into the contributions from when the same source is used as a cue ("same") and different source is used as a cue ("diff.").

strength effect. However, it was unclear as to whether this was due to increased interference from the source cues in the source memory task or due to a decision phenomenon where unrecognized items elicit guess responses (e.g.; Hautus et al., 2008). Thus, in Experiment 2 we implemented a conditionalized source memory procedure where source memory was only tested on recognized items. This experiment resulted in a null list strength effect, which makes it very unlikely that the list strength effect observed in Experiment 1 was due to extra interference from the source cues. In Experiment 3, we tested item recognition and source memory in separate test lists and collected confidence ratings. As the analysis of the source memory data was progressively restricted on confidence, we found stronger evidence for a null list strength effect in source memory, with the d' in the pure weak and mixed list condition being almost identical when the analysis was restricted to items that received a high confidence item recognition judgment. These results strongly suggest that a major constraint on models of item recognition memory, the null list strength effect, additionally applies to source memory. They coincide with other manipulations, such as word frequency, study-test delay, depth of processing, and aging, which appear to affect item

recognition and source memory in similar ways (Glanzer et al., 2004).

Tests for the presence of a list strength effect in source memory are critical, as they constrain the development of a mechanistic model of source memory. As it currently stands, the vast majority of source memory models' predictions are restricted to the shapes of the ROC functions in both item recognition and source memory, and are often mute with respect to predictions about variables and manipulations that are often of interest to theories of episodic memory. Specifically, an SDT model can accommodate a null list strength effect by assuming the d' parameter is unaffected by list strength, but it does not provide a principled explanation as to why this occurred. For this reason, we developed a model that is an extension of the Osth and Dennis (2015) global matching model, by assuming that participants form a conjunctive binding of the item, list context, and source. At retrieval, the item and list context are combined with each source cue and matched against the contents of memory. This is done for each source, and the difference between the memory strengths is used to drive a decision.

The model was applied to all experiments simultaneously using hierarchical Bayesian estimation and provided a good account of the data. We found evidence consistent with the hypothesis that the list strength effect observed in Experiment 1 was a decision level phenomenon by implementing a mixture model where only items that are recognized employ the source cues, otherwise a guess response is elicited, a mechanism inspired by the Hautus et al. (2008) model. We found this mixture model significantly improved on the original model, and was able to capture the list strength effect in the first experiment while also capturing the null effects of list strength in Experiment 2 and 3.

The model was also able to capture all of the other trends in the data, such as the decreased HR and FAR in the mixed list in item recognition, the word frequency effect in both item recognition and source memory, and the ROC shapes across tasks and conditions. The coverage of the trends is impressive when one considers that the only parameter that varied across the pure weak and mixed lists was the learning rate for strong

items; all other parameters were fixed and the predictions for the mixed list were derived from the structure of the model itself. The model predicted null list strength effects in both tasks for the same reason; the bulk of the interference contributions appear to come from pre-experimental sources rather than the experimental context itself, similar to previous work (Osth & Dennis, 2015).

The mechanism from the Hautus et al. (2008) model, where guesses are elicited on decisions where items are not recognized, was inspired by analyses from ROC data, which demonstrated that source memory abruptly dropped to chance when performance was restricted to unrecognized items. Other investigations have found source memory to be inaccurate for unrecognized items (Bell et al., 2017; Malejka & Broder, 2016). In our Experiment 3, an analysis of unrecognized items using a hierarchical Bayesian SDT model found that source memory performance for unrecognized items was slightly above chance but considerably lower than the accuracy for recognized items. Although this might appear to be contrary to the mechanism of the Hautus et al. (2008) model, our Experiment 3 used separate test phases for item recognition and source memory. If there is some degree of trial-to-trial criterion variability (e.g.; Benjamin et al., 2009), it's possible that some items which initially elicited a "no" response in the item recognition due to a high criterion were more likely to be recognized during the source test, eliciting a source retrieval instead of a guess, and so slightly greater than chance performance.

REM Model Predictions for List Strength Effects in Source Memory

So far, we have focused more on an extension of the Osth and Dennis (2015) model to source memory. Here, we discuss a couple different possible extensions to source memory of another prominent model of item recognition, REM (Shiffrin & Steyvers, 1997). In REM, items are represented as vectors with features sampled from a geometric distribution. By convention, the number of features in each vector is set to 20. During learning, a memory trace vector is created in memory. Features are copied into the trace vector from the item

vector with probability u; these sampled features are copied correctly into memory with probability c, otherwise, a new features is sampled with probability $1 - c$. During retrieval, a probe cue is matched against each trace in memory and a likelihood ratio is calculated reflecting the probability that the trace is the same as the probe divided by the probability that the trace is not the probe. These likelihood ratios are averaged across all traces in memory, and if the averaged likelihood ratio exceeds a decision criterion a "yes" response is made.

In Shiffrin and Steyvers (1997), associations are represented as concatenations among the to-be-associated items. That is, if an association between "cat" and "dog" is learned, a vector of 20 elements corresponding to "cat" is concatenated with a vector of 20 elements corresponding to "dog" to create a vector with 40 elements that corresponds to the association of the two vectors. A similar approach can be considered for source memory, where vectors corresponding to source A and source B are concatenated to each of the items; these item-source vectors are then learned according to the process described above. A source memory decision can be made by combining the probe item vector with the source A vector, cuing memory to assess the strength of source A, then repeating the process for the source B vector, and calculating the difference and comparing it to a decision criterion.

We generated simulations of the model for our paradigm by using LF and HF words. We followed conventional parameterizations of the model and used $c = .7$, $g = .3$ for LF words, and $g = .45$ for HF words, while using $g = .4$ for source vectors. We also used three different learning rates for weak items ($u = .16, .24, .32$). The learning rate for the strong items was manipulated across five levels, where the lowest learning rate was identical to the learning rate for weak items (this case ends up being equivalent to a pure weak list, as all items have the same learning rate).

Predictions can be seen in the left panel of Figure 9, where weak item predictions are shown in green and strong item predictions are shown in red. One can see that as the learning rate for strong items is increased, performance on the weak items is decreased.

There is a strong contrast to the predictions for item recognition, where it has been shown that the model predicts a null list strength effect across a range of different parameter values (Criss, 2006; Shiffrin & Steyvers, 1997). This is due to the effects of differentiation; as traces in memory are strengthened, their similarity to probe cues from other items decreases, effectively decreasing item interference as list strength is increased (Criss, 2006). However, Criss (2006) demonstrated that the predictions of differentiation are specific to the case where items are dissimilar to each other. In the source memory case, half of the traces in memory will share as many as 50% of their features with the probe due to the matching source features. Criss (2006) demonstrated that under such conditions, traces with 50% similarity to the probe will actually show increasing interference as their strength is increased.

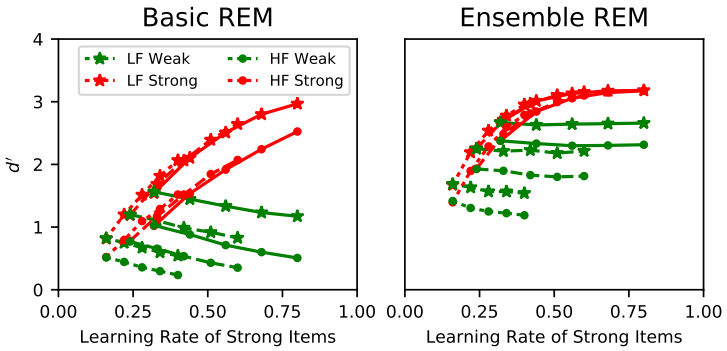

Figure 9. REM model predictions for our list strength paradigm in source memory. The left panel shows predictions from the simple concatenation model proposed by Shiffrin and Steyvers (1997) while the right panel shows predictions from the Criss and Shiffrin (2005) ensemble features model. See the text for details of the simulations.

However, very different predictions result from when ensemble features are employed. Criss and Shiffrin (2005) proposed an extended REM model where each concatenated vector additionally includes a set of ensemble features that are unique to the combination of the features. That is, if the word "truck" is combined with a source A vector, an additional vector is additionally concatenated to this vector that contains features that are

unique to the ensemble. That is, if another word such as "metal" is combined with a source A vector, its ensemble features will be different from the ensemble features that correspond to the truck + source A pairing.

We simulated performance from the model with ensemble features where each ensemble vector constituted 20 features and was sampled with $g = .4$. This led to the storage of concatenated item, source, and ensemble vectors with a total of 60 features. Predictions from the ensemble REM model can be seen in the right panel of Figure 9. It is immediately evident that the performance of the weak items is effectively flat as the learning rate of the strong items is increased over a very wide range. Essentially, a null list strength effect is predicted, in accordance with our data.

In the ensemble model, although half of the traces will still bear a resemblance to the probe due to the matching source features, the number of expected shared features is only 33% instead of 50% due to the presence of the unique ensemble features. As the similarity between a probe vector and a trace vector is reduced, differentiation is more likely to reduce interference as strength is increased, which will lead to the prediction of a null list strength effect. The functional explanation is quite similar to the explanation from our model, in that both models claim that increases in list strength do not significantly increase interference in source memory. Nonetheless, the models differ in that the REM model claims the lack of interference is due to the process of differentiation, while our model claims the lack of interference is due to the bulk of interference coming from memories acquired prior to the experiment. We did not competitively test the models in our experiment because the simulation intensive nature of REM makes it very difficult to apply it to data using the methods we have employed.

Conclusions

The finding of a null list strength effect in item recognition was very influential, initiating the development of a whole new set of memory models. Here, we extended the

list strength paradigm to source memory across three experiments varying in the nature of the test phase. Our results suggest that list strength manipulations do not increase interference in source memory. This finding was reinforced by fitting an extension of the Osth and Dennis (2015) global matching model, which was capable of jointly addressing all aspects of the item recognition and source memory data in each experiment.

References

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267–273.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*(4), 548–564.

Bell, R., Mieth, L., & Buchner, A. (2017). Emotional memory: No source memory without old-new recognition. *Emotion*, *17*(1), 120–130.

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

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608.

Buratto, L. G., & Lamberts, K. (2008). List strength effect without list length effect in recognition memory. *Quarterly Journal of Experimental Psychology*, *61*(2), 218–226.

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.

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., 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. (2005). List discrimination in associative recognition and implications for representation. *Journal of Experimental Psychology: Leanring, Memory, and Cognition*, *31*(6), 1199–1212.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65-70.

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. (2003). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 767–778.

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.

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

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., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1176–1195.

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

Guttentag, R. E., & Carroll, D. (1994). Identifying the basis for the word frequency effect in recognition memory. *Memory*, *2*, 255–273.

Guttentag, R. E., & Carroll, D. (1997). Recollection-based recognition: Word frequency effects. *Journal of Memory and Language*, *37*, 502–516.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review*, *15*(5), 889–905.

Heathcote, A., Raymond, F., & Dunn, J. (2006, November). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*(4), 495–514.

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.

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.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

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.

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

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*(4), 465–478.

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.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models.

Journal of Mathematical Psychology, *55*, 1–7.

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

Malejka, S., & Broder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition*, *44*(1), 63–72.

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.

Marsh, R. L., Cook, G. I., & Hicks, J. L. (2006). The effect of context variability on source memory. *Memory & Cognition*, *34*(8), 1578–1586.

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.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology*, *10*, 465–501.

Mulligan, N. W., & Osborn, K. (2009). The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 564–571.

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

Murdock, B. B. (2006). Decision-making models of remember-know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, *113*(3), 648–656.

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.

Norman, K. A. (2002). Differential Effects of List Strength on Recollection and Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1083–1094.

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

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

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

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

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

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

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

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., 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.

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.

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

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., 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.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, *33*, 151–170.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition*, *36*(1), 1–8.

Starns, J. J., & Ksander, J. C. (2016). Item strength influences source confidence and alters source memory zROC slopes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 351–365.

Starns, J. J., Pazzaglia, A. M., Rotello, C. M., Hautus, M. J., & Macmillan, N. A. (2013). Unequal-strength source zROC slopes reflect criteria placement and not (necessarily) memory processes. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*, 1377–1392.

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.

Stretch, V., & Wixted, J. T. (1998). 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.

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

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.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

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

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1415–1434.

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.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characterics (ROCs) in recognition memory: A review. *Psychological Butten*, *133*(5), 800–832.

Appendix A

Analytic Derivation of the Source Memory Model

Following Humphreys, Pike, Bain, and Tehan (1989), we can deconstruct Equation 2 into the various components that comprise the memory tensor M. We are writing this for the case where source A is used as a cue, but the math equally applies for the source B case. While subscripts a and b denote source A and B in the experiment, we use subscript c to refer to other sources acquired before the experiment:

$$
\begin{aligned}
s = (I'_t \otimes C'_s \otimes S'_a).[r(I_t \otimes C_s \otimes S_a) & && \text{Self Match} \\
&+ \sum_{i \in A, i \neq t} r(I_i \otimes C_s \otimes S_a) && \text{Item Noise (A)} \\
&+ \sum_{i \in B, i \neq t} r(I_i \otimes C_s \otimes S_b) && \text{Item Noise (B)} \\
&+ \sum_{u \in P, u \neq s} (I_t \otimes C_u \otimes S_a) && \text{Context Noise (A)} \\
&+ \sum_{u \in P, u \neq s} (I_t \otimes C_u \otimes S_b) && \text{Context Noise (B)} \\
&+ \sum_{u \in P, u \neq s} (I_t \otimes C_u \otimes S_c) && \text{Context Noise (C)} \\
&+ \sum_{u \in P, u \neq s, z \notin L} (I_z \otimes C_u \otimes S_a) && \text{Background Noise (A)} \\
&+ \sum_{u \in P, u \neq s, z \notin L} (I_z \otimes C_u \otimes S_b) && \text{Background Noise (B)} \\
&+ \sum_{u \in P, u \neq s, z \notin L} (I_z \otimes C_u \otimes S_c)] && \text{Background Noise (C)}
\end{aligned}
\tag{3}
$$

where i indicates items on the study list that are not the item cue, which is referred to by subscript t. A and B denote the sets of items that were studied in source A and B, respectively. The u subscript indicates a prior list context from the set of all contexts prior to the study list (P), and z indicates items from prior list contexts that were not on the study list.

In linear algebra, the dot product of two outer products $((A \otimes B)\dot{(}C \otimes D))$ is equal to

the product of the dot products of the constituent vectors $((A \cdot C)(B \cdot D)$. Using that, we can rewrite Equation 3 as the match between the cue vectors and the stored vectors:

$$
\begin{aligned}
s = & r(I'_t.I_t)(C'_s.C_s)(S'_a.S_a) + && \text{Self Match} \\
& \sum_{i \in A, i \neq t} r(I'_t.I_i)(C'_s.C_s)(S'_a.S_a) + && \text{Item Noise} \\
& \sum_{i \in B, i \neq t} r(I'_t.I_i)(C'_s.C_s)(S'_a.S_b) + && \text{Item Noise} \\
& \sum_{u \in P, u \neq s} (I'_t.I_t)(C'_s.C_u)(S'_a.S_a) + && \text{Context Noise} \\
& \sum_{u \in P, u \neq s} (I'_t.I_t)(C'_s.C_u)(S'_a.S_b) + && \text{Context Noise} \\
& \sum_{u \in P, u \neq s} (I'_t.I_t)(C'_s.C_u)(S'_a.S_c) + && \text{Context Noise} \\
& \sum_{u \in P, u \neq s, z \notin L} (I'_t.I_z)(C'_s.C_u)(S'_a.S_a) + && \text{Background Noise} \\
& \sum_{u \in P, u \neq s, z \notin L} (I'_t.I_z)(C'_s.C_u)(S'_a.S_b) + && \text{Background Noise} \\
& \sum_{u \in P, u \neq s, z \notin L} (I'_t.I_z)(C'_s.C_u)(S'_a.S_c) && \text{Background Noise}
\end{aligned}
\tag{4}
$$

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

$$
\begin{aligned}
C'_s.C_s &\sim Normal(\mu_{ss}, \sigma^2_{ss}) && \text{Context Match} && (5)\\
C'_s.C_u &\sim Normal(0, \sigma^2_{su}) && \text{Context Mismatch} &&\\
I'_t.I_t &\sim Normal(\mu_{tt}, \sigma^2_{tt}) && \text{Item Match} &&\\
I'_t.I_i &\sim Normal(0, \sigma^2_{ti}) && \text{Item Mismatch} &&\\
S'_a.S_a &\sim Normal(\mu_{aa}, \sigma^2_{aa}) && \text{Source Match} && (6)\\
S'_a.S_b &\sim Normal(0, \sigma^2_{ab}) && \text{Source Mismatch} &&
\end{aligned}
$$

(7)

The means and variances of the distributions of dot products are the parameters of the model, although as we note in Appendix B several of these parameters are fixed to improve the estimation of the remaining parameters. This approach is similar to the kernel trick employed by support vector machines (Schölkopf & Smola, 2002). The choice of the normal distribution offers mathematical convenience for this application by allowing separate specification of the mean and variance parameters. Covariances were avoided by fixing the means of the mismatch distributions to zero. Other parameters of the model were fixed to reduce the number of free parameters and because they were not found to be critical for the performance of the model.

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

$$
\begin{aligned}
E(X_1, ..., X_n) &= \prod_i E(X_i)\\
V(X_1, ..., X_n) &= \prod_i (var(X_i) + E(X_i)^2) - \prod_i E(X_i)^2)
\end{aligned}
$$

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 of an item studied in source A when source A ($\mu_{a|a}$) is used as a cue is simply the learning rate r, while the mean of the source A distribution when B is used as a cue ($\mu_{b|a}$) is zero. The variances are:

$$\mu_{a|a} = r\mu_{tt}\mu_{ss}\mu_{aa} \tag{8}$$

$$\mu_{b|a} = 0 \tag{9}$$

$$\begin{aligned}
\sigma^2_{a|a} =& r^2[(\sigma^2_{tt}+\mu^2_{tt})(\sigma^2_{ss}+\mu^2_{ss})(\sigma^2_{aa}+\mu^2_{aa}) - (\mu^2_{tt}\mu^2_{ss}\mu^2_{aa}) && \text{Self Match} \\
& r^2(l/2-1)[\sigma^2_{ti}(\sigma^2_{ss}+\mu^2_{ss})(\sigma^2_{aa}+\mu^2_{aa})] && \text{Item Noise} \\
& l/2r^2[\sigma^2_{ti}\sigma^2_{ab}(\sigma^2_{ss}+\sigma^2_{ss})]+ && \text{Item Noise} \\
& n_a[\sigma^2_{su}(\sigma^2_{tt}+\mu^2_{tt})(\sigma^2_{aa}+\mu^2_{aa})]+ && \text{Context Noise} \\
& n_b[\sigma^2_{su}(\sigma^2_{tt}\sigma^2_{ab})]+ && \text{Context Noise} \\
& n_c[\sigma^2_{su}(\sigma^2_{tt}\sigma^2_{ac})]+ && \text{Context Noise} \\
& m_a[(\sigma^2_{ti}\sigma^2_{su})(\sigma^2_{aa}+\mu^2_{aa})]+ && \text{Background Noise} \\
& m_b(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ab})+ && \text{Background Noise} \\
& m_c(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ac}) && \text{Background Noise}
\end{aligned} \tag{10}$$

$$
\begin{aligned}
\sigma^2_{b|a} =& r^2[\sigma^2_{ba}(\sigma^2_{tt}+\mu^2_{tt})(\sigma^2_{ss}+\mu^2_{ss})]+ && \text{Self Match} \\
& r^2(l/2-1)[\sigma^2_{ti}\sigma^2_{ba}(\sigma^2_{ss}+\mu^2_{ss})]+ && \text{Item Noise} \\
& l/2r^2[\sigma^2_{ti}\sigma^2_{bb}(\sigma^2_{ss}+\mu^2_{ss})]+ && \text{Item Noise} \\
& n_a[\sigma^2_{su}\sigma^2_{ba}(\sigma^2_{tt}+\mu^2_{tt})]+ && \text{Context Noise} \\
& n_b[\sigma^2_{su}(\sigma^2_{tt}+\mu^2_{tt})(\sigma^2_{bb}+\mu^2_{bb})]+ && \text{Context Noise} \\
& n_c]\sigma^2_{su}\sigma^2_{bc}(\sigma^2_{tt}+\mu^2_{tt})]+ && \text{Context Noise} \\
& m_a(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ba})+ && \text{Background Noise} \\
& m_b[(\sigma^2_{ti}\sigma^2_{su}(\sigma^2_{bb}+\mu^2_{bb})]+ && \text{Background Noise} \\
& m_c(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{bc}) && \text{Background Noise}
\end{aligned} \tag{11}
$$

Because we expect symmetry between the two sources (equal performance between source A and source B), we assume parameters that correspond to when source B was studied are the same as those as when source was studied. More specifically, we assume $\mu_{bb} = \mu_{aa}$, $\sigma^2_{bb} = \sigma^2_{aa}$, $\sigma^2_{ba} = \sigma^2_{ab}$, and $\sigma^2_{bc} = \sigma^2_{ac}$. With these assumptions, the expressions above can be rewritten for the $a|b$ and $b|b$ cases by substituting μ_{aa} with μ_{bb}, σ^2_{bb} with σ^2_{aa}, σ^2_{ab} with σ^2_{ba}, and σ^2_{ac} with σ^2_{bc}.

To arrive at memory strength distributions for source A and source B, we can take the difference between the source A and source B cues. For source A, this involves the difference between the $a|a$ and $b|a$ distributions while source B involves the difference between the $a|b$ and $b|b$ distributions. The mean of the difference between two normal distributions y and z is

$$
\begin{aligned}
\mu_{y-z} =& \mu_y - \mu_z \\
\sigma^2_{y-z} =& \sigma^2_y + \sigma^2_z - cov_{y-z}
\end{aligned}
$$

It may seem at first glance that there would be a correlation between the memory

strengths of the different source cues due to the re-usage of the item and context cues for each distribution. However, the item and context cues are multiplied by the source cues. We performed simulations and found that the correlation between the products of three normal distributions with two overlapping distributions is zero when the mean of one of those products is zero. For example, consider four normal distributions, $a\ Normal(1,1)$, $b\ Normal(1,1)$, $c\ Normal(1,1)$, and $d(Normal(0,1)$. Simulations demonstrated that the correlation between $a*b*c$ and $a*b*d$ is approximately zero. We were not able to demonstrate this analytically because to our knowledge there are no analytics available for the products of normal distributions when the components of the products are correlated. Given that we can safely assume the covariances to be zero:

$$\sigma_A^2 = \sigma_{a|a}^2 + \sigma_{a|b}^2 \tag{12}$$

$$\sigma_B^2 = \sigma_{b|a}^2 + \sigma_{b|b}^2 \tag{13}$$

As mentioned in the main text, to derive predictions about item recognition, we assume that in place of the source cue, participants employ a generalized cue that matches each source vector in memory with a strength of one and no variance. This has the effect of collapsing across the background memories ($n_{item} = n_a + n_b + n_c$ and $m_{item} = m_a + m_b + m_c$) as the different source vectors studied with each prior memory to not influence the resulting memory strength. This produces the following expressions:

$$\mu_{old} = r\mu_{tt}\mu_{ss} \tag{14}$$

$$\mu_{new} = 0 \tag{15}$$

$$\begin{aligned}
\sigma^2_{old} =& r^2[(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{ss} + \mu^2_{ss}) - (\mu^2_{tt}\mu^2_{ss}) && \text{Self Match} \\
& r^2(l-1)[\sigma^2_{ti}(\sigma^2_{ss} + \mu^2_{ss})] && \text{Item Noise} \\
& n_{item}[\sigma^2_{su}(\sigma^2_{tt} + \mu^2_{tt})]+ && \text{Context Noise} \\
& m_{item}(\sigma^2_{ti}\sigma^2_{su}) && \text{Background Noise}
\end{aligned} \tag{16}$$

$$\begin{aligned}
\sigma^2_{new} =& r^2 l[\sigma^2_{ti}(\sigma^2_{ss} + \mu^2_{ss})] && \text{Item Noise} \\
& n_{item}[\sigma^2_{su}(\sigma^2_{tt} + \mu^2_{tt})]+ && \text{Context Noise} \\
& m_{item}(\sigma^2_{ti}\sigma^2_{su}) && \text{Background Noise}
\end{aligned} \tag{17}$$

Figure A1 compares the distributions produced by the analytic approximation to simulations for both item recognition (left panel) and source memory (right panel). The model was simulated by drawing 500,000 samples from normal distributions and combining them via Equation 4. We found a very strong correspondence between the analytic approximation and the simulations. These simulation results also demonstrate the normal approximation for sums of products of normal distributions is a reasonable description of the distribution when large numbers of products of normal distributions are summed together.

The likelihood ratio transformation

As mentioned in the main text, in order to capture the mirror effect in item recognition, we apply the memory strengths described above to a log likelihood ratio transformation to capture the mirror effect using the linear approximation developed by Osth, Dennis, and Heathcote (2017) which results in normally distributed log likelihood ratios, which we denote using λ. These expressions were written for the general case in terms of discrimination d and the relative variability of the target distribution S, which we can reach by normalizing the parameters by σ_{new}:

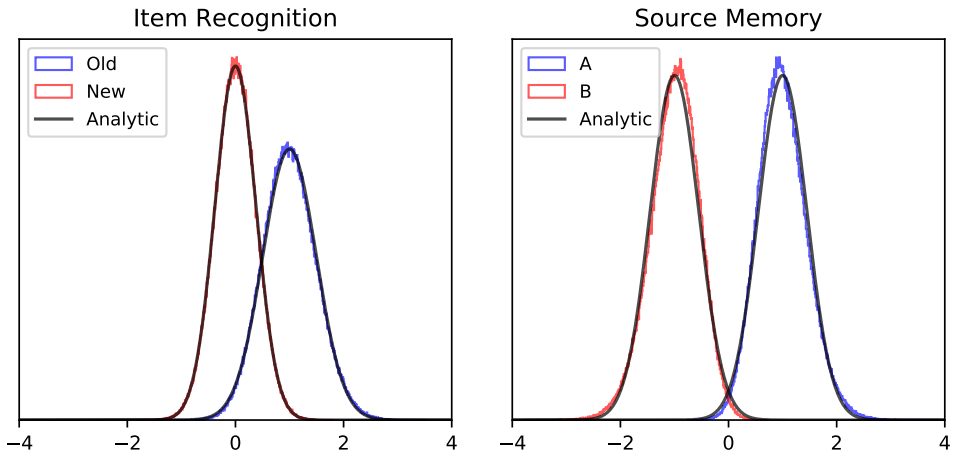

Figure A1. Histograms of simulation predictions along with analytic approximations (lines) for the item recognition (left) and source memory (right) model. Model parameters were $r, \mu_{tt}, \mu_{ss}, \mu_{aa} = 1$, $\sigma^2_{tt}, \sigma^2_{ss} = .05$, $\sigma^2_{aa}, \sigma^2_{ab} = .01$, $\sigma^2_{ti} = .001$, $n_{aa}, n_{ab} = 2$, $n_{ac} = 6$, $m_{aa}, m_{ab} = 10$, $m_{ac} = 480$.

$$d_{item} = \mu_{old}/\sigma_{new} \tag{18}$$

$$S_{item} = \sigma_{old}/\sigma_{new} \tag{19}$$

For source memory, the variances of both distributions are equal, so we can divide by the variability of either the A or B distribution:

$$d_{source} = (\mu_A - \mu_B)/\sigma_A \tag{20}$$

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

$$\mu_{\lambda new} = -((\frac{d^2}{2})(\frac{S^2+3}{4S^2}) + log(S)) \tag{21}$$

$$\mu_{\lambda old} = d^2\frac{S^2+1}{2S^2} + \mu_{\lambda new} \tag{22}$$

$$\sigma_{\lambda new} = d\frac{S^2+1}{2S^2} \tag{23}$$

$$\sigma_{\lambda old} = S\sigma_{\lambda L} \tag{24}$$

For source memory, given that the two distributions have equal variance, we can follow the expressions of Glanzer et al. (2009):

$$\mu_{\lambda A} = d^2/2 \tag{25}$$

$$\mu_{\lambda B} = -d^2/2 \tag{26}$$

$$\sigma_{\lambda A,B} = d^2 \tag{27}$$

In mixed lists of weak and strong items, using the above expressions imply that the participants know whether an item is weak or strong before having seen the item. In these cases, we subject the true memory strengths to an expected distribution that is the average of the weak and strong items (e.g.; Osth & Dennis, 2015; Starns et al., 2010). This can be accomplished by averaging the learning rates from the two strength conditions to generate r_{avg} and then generating the expected strengths d and S according to the above equations. The actual learning rates r_{weak} and r_{strong} are used to generate the true strengths for a given condition, which we denote as d^* and S^*. Expressions for the target distributions of a mixed strength list in item recognition are thus:

$$\mu_{\lambda old} = dd^*\frac{S^2+1}{2S^2} + \mu_{\lambda L} \tag{28}$$

$$\sigma_{\lambda old} = S^*\sigma_{\lambda L} \tag{29}$$

The lure expressions for a mixed list are unchanged. For source memory, we have:

$$\mu_{\lambda A} = d * d/2 \tag{30}$$

$$\mu_{\lambda B} = -d * d/2 \tag{31}$$

$$\sigma_{\lambda A,B} = d \tag{32}$$

Appendix B

Prior Distributions on Model Parameters

Several parameters of the model were fixed to improve estimation of the remaining parameters and because they were not found to greatly contribute to the fit of the model when they were freely estimated. The self match variability parameters for items (σ_{tt}^2) and context (σ_{ss}^2) govern the ratio of target-to-lure variability in the model. However, we found in practice that we were able to yield good fits to the ROC function by only varying one of those parameters. We fixed σ_{ss}^2 to .1, which was the mean of the group level distribution found by Osth and Dennis (2015). In addition, the means of match distributions for items (μ_{tt}), contexts (μ_{ss}) and sources (μ_{aa}) were all fixed to one. It would be possible to estimate these parameters if the strengths of each of these dimensions were manipulated, via either stimulus strength, study-test delay, or source discriminability, but given that none of these manipulations were present we were able to achieve good fits by fixing each of these parameters.

We additionally fixed the number of prior memories for LF words, $n_{LF,item}$, to 20 and the total number of background memories, m_{item}, to 10e6, while freely estimating the prior occurrences of HF words, $n_{HF,item}$. We fixed these parameters because it is not possible to identify both the number and strength of the prior memories. The values we chose were arbitrary, and other values yielded similar results. We similarly fixed the prior occurrences of items in each source, $m_{aa}, m_{ab}, n_{aa}, n_{ab}$, to 5% of the total memories (e.g.; $n_{LF,aa} = .05 n_{LF,item}$), which leaves the number of memories in non-studied sources ($n_{LF,ac}, n_{HF,ac}, m_{ac}$) as 90% of the total number of prior memories. We initially estimated the proportion of prior memories that match the sources in the experiment as a free parameter, but this did not greatly improve the fit of the model.

All parameters that were bounded from zero onward were sampled on a log scale, which allows for sampling from a normal prior distribution. Subject level parameters were sampled from group level distributions with mean M and standard deviation ς:

$$log(\sigma^2_{ti}) \sim Normal(M_{\sigma ti}, \varsigma_{\sigma ti})$$

$$log(\sigma^2_{tt}) \sim Normal(M_{\sigma tt}, \varsigma_{\sigma tt})$$

$$log(\sigma^2_{su}) \sim Normal(M_{\sigma su}, \varsigma_{\sigma su})$$

$$log(\sigma^2_{aa}) \sim Normal(M_{\sigma aa}, \varsigma_{\sigma aa})$$

$$log(\sigma^2_{ab}) \sim Normal(M_{\sigma ab}, \varsigma_{\sigma ab})$$

$$log(\sigma^2_{ac}) \sim Normal(M_{\sigma ac}, \varsigma_{\sigma ac})$$

$$log(n_{HF,item}) \sim Normal(M_{nHFitem}, \varsigma_{nHFitem})$$

$$log(r_{weak,j}) \sim Normal(M_{rweakj}, \varsigma_{rweakj})$$

where j is the experiment (1, 2, or 3). The learning rates for strong items in Experiment j were determined as:

$$r_{strong,i} = r_{weak,i} * (1 + rs) \tag{33}$$

where rs is a scalar on the $(0, \infty)$ interval. One is added to rs to ensure that the learning rates for strong items cannot be weaker than the learning rates for weak items. rs is also sampled on a log scale, but unlike the r_{weak} parameters, it does not vary across experiments:

$$log(rs) \sim Normal(M_{rs}, \varsigma_{rs})$$

Item and source criteria were sampled from normal distributions, along with the d' parameters in the hierarchical SDT models used in the analysis of Experiment 3:

$$
\begin{aligned}
d' \sim & Normal(M_d, \varsigma_d) \\
\phi_{k,j} \sim & Normal(M_{\phi kj}, \varsigma_{\phi kj} \\
\phi_{k,3,3} \sim & Normal(M_{\phi t,3,3}, \varsigma_{\phi t,3,3}
\end{aligned}
\tag{34}
$$

where k refers to the task (item vs. source) and j refers to the experiment (1 or 2). $\phi_{k,3,3}$ is the central criterion of the five criteria for Experiment 3. The remaining criteria were determined relative to the central criterion as:

$$\phi_{k,1} = \phi_{k,3} - c_{k,1} \tag{35}$$

$$\phi_{k,2} = \phi_{k,3} - c_{k,2} \tag{36}$$

$$\phi_{k,4} = \phi_{k,3} + c_{k,4} \tag{37}$$

$$\phi_{k,5} = \phi_{k,3} + c_{k,5} \tag{38}$$

where the lower (liberal) criteria are determined using lower numbers (1 and 2), and the higher (conservative) criteria are denoted using higher numbers (4 and 5). This parameterization was done to improve sampling as it guaranteed a partial ordering of the decision criteria. The c parameters were sampled from truncated normal (TN) distributions that were truncated from zero to infinity:

$$
\begin{aligned}
c_{k,1} \sim & TN(M_{citem1}, \varsigma_{ck1}, 0, \infty) \\
c_{k,2} \sim & TN(M_{citem2}, \varsigma_{ck2}, 0, \infty) \\
c_{k,4} \sim & TN(M_{citem4}, \varsigma_{ck4}, 0, \infty) \\
c_{k,5} \sim & TN(M_{citem5}, \varsigma_{ck5}, 0, \infty)
\end{aligned}
\tag{39}
$$

Priors on the group level distributions for the majority of the parameters were relatively non-informative:

$$\begin{aligned} M_{\sigma ti,\sigma tt,\sigma su,\sigma aa,\sigma ab,\sigma ac,nHFitem,rweaki,d} &\sim Normal(0, 10) \\ M_{\phi ti} &\sim Normal(0, 1) \\ M_{ckj} &\sim TN(.5, .5, 0, \infty) \\ \varsigma_{\sigma ti,\sigma su,\sigma aa,\sigma ab,\sigma ac,nHFitem,rweakj,\phi ti,ckj,d} &\sim Gamma(1, 3) \end{aligned}$$

We adopted somewhat stricter priors for the rs and σ^2_{tt} parameters. This was because each parameter was only partially constrained across the three experiments. In the case of the rs parameter, Experiment 3 did not include tests of strong items; thus, this parameter is primarily estimated from Experiments 1 and 2. In addition, only Experiment 3 contained ROC functions, which constrains the estimates of σ^2_{tt}. To improve estimation of these parameters, we used stricter prior distributions on ς, which place much higher likelihoods on lower values of ς:

$$\begin{aligned} M_{rs} &\sim Normal(0, .5) \\ \varsigma_{rs,\sigma tt} &\sim Gamma(1, 25) \end{aligned} \tag{40}$$