Journal of Memory and Language 103 (2018) 91–113

Contents lists available at ScienceDirect

Journal of Memory and Language

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

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

Adam F. Osth[a,*], Julian Fox[a], Meredith McKague[a], Andrew Heathcote[b], Simon Dennis[a]

[a] The University of Melbourne, Australia
[b] The University of Tasmania, Australia

ARTICLE INFO

Keywords:
Recognition memory
Source memory
Global matching models
List strength effect

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.

Introduction

A distinction in episodic memory concerns the difference between information about learned content and the context in which it occurred. A common memory failure is when one remembers a fact or detail but has no memory for where he or she 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; Glanzer, Hilford, & Kim, 2004; Hautus, Macmillan, & Rotello, 2008; Klauer & Kellen, 2010; Slotnick & Dodson, 2005; Yonelinas, 1999). 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).

While such models yield useful predictions about the shapes of the receiver operating characteristic (ROC) in each task (Slotnick & Dodson, 2005) and whether source memory is accurate without item memory (Starns, Hicks, Brown, & Martin, 2008), 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. This is because these models define the form of the decision variable but are agnostic as to the encoding, storage, and retrieval assumptions that give rise to it. In contrast, the class of global matching models has made such specifications (Clark & Gronlund, 1996). In global matching models, memory strength is determined by the similarity between the retrieval cues and each stored item in memory; these similarities are summed (or averaged; Shiffrin & Steyvers, 1997) to produce a single strength value that can be compared to a response

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

https://doi.org/10.1016/j.jml.2018.08.002
Received 16 January 2018; Received in revised form 31 July 2018
Available online 17 August 2018

criterion to make a decision. Collectively, the current generation of episodic recognition 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).

Nonetheless, many recent mechanistic models in the episodic memory literature have often been restricted to a single task and have rarely provided joint accounts of multiple memory tasks (but see Lehman & Malmberg, 2013, for a noteworthy exception). Hintzman (2011) criticized this tendency and argued that this has been leading to limited conclusions about the nature of memory as a whole. Consistent with this criticism, current mechanistic models of recognition memory have experienced little, if any, extension to the source memory paradigm. The current article attempts to fill this gap by testing one of the major constraints of episodic memory models, the list strength effect (LSE), in a source memory paradigm, and further introduce an extension of the Osth and Dennis (2015) model to provide a joint account of the results from both item recognition and source memory. The list strength paradigm asks the question *can strengthening a memory cause forgetting of other memories?*

The list strength paradigm: data and model predictions

A prediction of global matching models is that as the number of items in memory is increased, performance should decrease. In these 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; repetitions are treated in the same manner as increases in the number of studied items and contribute additional noise at retrieval.

To understand how this prediction is manifested, consider a sequence of study list items such as ABCD. Most models would predict that strengthening A and B via study time and/or repetition should increase performance on A and B. The counter-intuitive prediction that emerged from these models is that strengthening A and B should *impair* performance on C and D. This prediction can be tested by comparing lists with different compositions of strengthened and non-strengthened items, such as a pure weak list where all items are presented once (ABCD) and a mixed list where half the items are presented once and half the items are presented four times (AAAABBBBCD). The original global matching models predicted that performance of the once presented items (C and D) should be worse in the mixed list than in the pure weak list due to the extra interference from the repetitions of A and B. The list with more repeated items would be considered a list with higher *list strength*.

A large 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 (Diana & Reder, 2005; Hirshman, 1995; 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 when strengthening is achieved via spaced presentations (Tulving & Hastie, 1972) but not when strengthening is achieved via massed presentations or depth of processing (Malmberg & Shiffrin, 2005). As a consequence of this failed prediction, amendments to the global matching framework were proposed that enabled the models to predict a null list strength effect (LSE) 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 than weak traces, whereas in the older models the opposite was the case.

Another class of models has argued that the null LSE is more indicative of interference stemming from sources other than the studied items (Dennis & Humphreys, 2001; Murdock & Kahana, 1993a, 1993b; Osth & Dennis, 2015). While initial models assumed that memory is a "blank slate" before presentation of the study list,[1] these models instead assume an interference contribution from pre-experimentally learned memories consisting of prior occurrences of the cue word (context noise) or from other memories in general (background noise). When such interference contributions are substantial, 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. A simple extension of these models to source memory would involve binding each item to 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. An example is depicted in Fig. 1, where "truck" and "joker" were studied in source A (red) and "sky" and "phone" in source B (green). 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 (s_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 (s_B). The difference between the memory strengths for source A and B could be used to make a decision - if this difference exceeds a decision criterion (ϕ_{source}) source A would be chosen, otherwise source B would be chosen.

Although this mechanism is similar to item recognition, the representational structure of the memory set in the source memory task can lead to different predictions. 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 source memory, half of the items in the list match the source cue, meaning that source memory can resemble cases where half of the representations in memory bear a high similarity to the retrieval cues.

We found this higher similarity in the task was sufficient to induce an LSE in the original version of the retrieving effectively from memory model (REM: Shiffrin & Steyvers, 1997); these simulations are detailed later in the General Discussion. 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 and produce a null LSE. However, differentiation only reduces interference when the similarity between the trace and the cue is relatively low. When the similarity is high, which is the case when 50% of traces match the source cue, interference increases with strength (Criss, 2006). However, later formulations of REM allow additional ensemble features that are unique to a binding between items or features (Criss & Shiffrin, 2005). Interestingly, ensemble features

[1] A reviewer pointed out that the central commitment of such models is not that there are no memories, but that any interference from prior memories is negligible. Another possibility is that interference from prior memories is eliminated because the context of the study list is sufficiently isolated from prior memories.

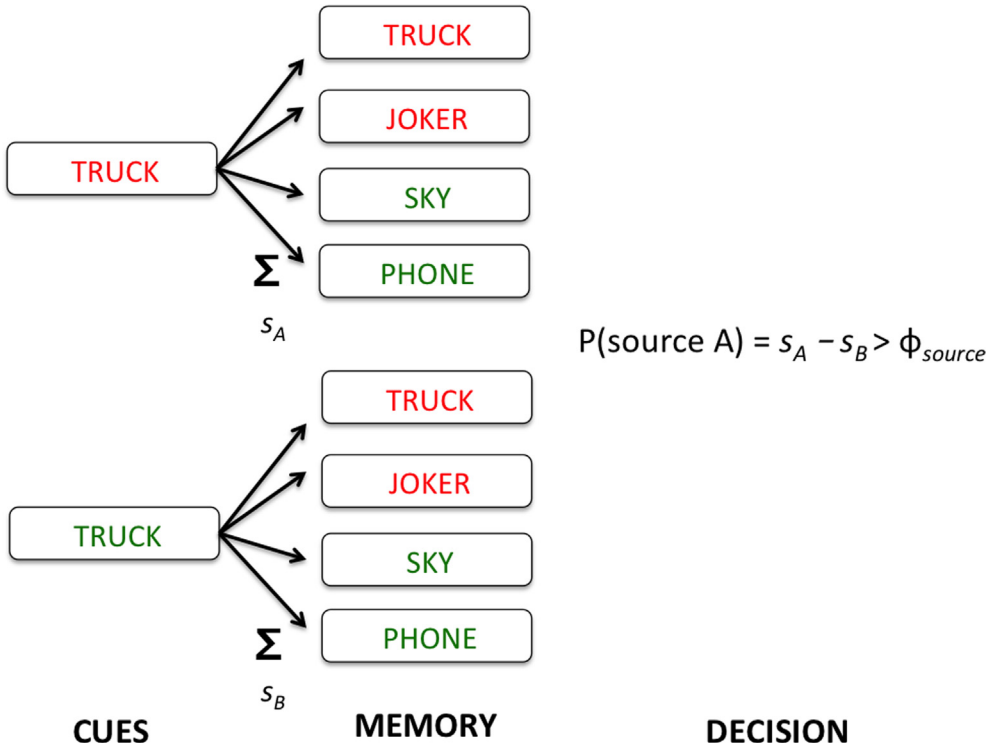

Fig. 1. A global matching account of the source memory task. The probe word "truck" is cued with the source A context (red, upper left) and matched against each of the study list items, each of which was bound to either the source A or source B (green) context, to measure the strength of the source A context s_A. In addition, "truck" is cued with the source B context to measure the strength of the source B context s_B. Source A is chosen if the difference between s_A and s_B exceeds a decision criterion Φ_{source}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mitigate the interfering effect of the matching source features and a null LSE is predicted.

The Osth and Dennis (2015) model makes predictions that closely align with the ensemble version of REM when interference among the items is minimal. In item recognition, Osth and Dennis demonstrated that predictions for list strength in the model depend on the magnitude of the interference between the studied items. As the interference between items approaches zero, the strength of the other memories cannot influence performance and no effect of list strength is predicted. Osth and Dennis accounted for the null LSE with word stimuli because the interference between words was relatively small.

In the source memory extension of the model, the similarity of each memory to the retrieval cues is a multiplication of the similarity to the item cue and the similarity to the source cue. Because of this multiplicative combination, when the similarity to the item cue is very low, the overall match from the memory is low even if its source information matches the source cue. As an example, consider if memory is cued with "truck in source A" in a list where "monkey" was studied in source A four times and "cup" was studied in source B once. As the similarity between the cue "truck" and each studied word ("monkey" and "cup") approaches zero, the overall match for each memory should approach zero even though "monkey" has a strong match to the source features. As the similarity between "truck" and the studied words increases, there should be more interference from the word "monkey" due to its stronger match on the item features. We have postponed a mathematical description of the model until later in the article, where we demonstrate that the model predicts a null LSE in both item recognition and source memory under a variety of different parameters when the interference among the items is very low, which is what would be expected here given that prior investigations with the model have found low interference among words (Osth & Dennis, 2015; Osth, Jansson, Dennis, & Heathcote, 2018).

There are additionally some well specified dual process models that predict a dissociation between item recognition and source memory with regards to the LSE. In dual process models, recognition decisions are based on either familiarity or recollection. In the source of activation confusion (SAC) model (Reder et al., 2000), the recollection component is composed of items being bound to an episode node. Items compete for activation from the episode node and strengthened items receive episodic activation at the expense of the weaker items on the list. Familiarity is a non-competitive baseline strength that is incremented upon study presentation and is thus unaffected by increases in list strength but contains no contextual or source information. In SAC, a null LSE is predicted in item recognition because the recollection deficit is compensated by relying more on familiarity in conditions of higher list strength, but an LSE is predicted in tasks that rely heavily on recollection, and source memory is one such example (Diana & Reder, 2005). In addition, the Norman and O'Reilly (2003) neural network model contains a familiarity-based cortical layer which is unaffected by list strength, as well as a hippocampal network that is necessary for tasks such as source memory or plurality discrimination that is impaired by increasing list strength. Both models have been argued to be consistent with the finding that an LSE is observed in the plurality discrimination task, where participants have to distinguish between studied items and switched plurality lures (i.e.: rejection the word "cat" if "cats" was studied), a task presumed to require recollection.

The current investigation

To our knowledge, the effects of list strength on source memory performance have not been investigated. A number of researchers have investigated 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) but 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). Participants rated words for pleasantness while viewing the words, which has been shown to increase performance in both item recognition and source memory tasks (Glanzer et al., 2004). In the pure weak (PW) condition each item was presented once (1×) while in the mixed condition half the items were presented once and the other half were presented four times (4×) with each repetition spaced throughout the study list. 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.

Unlike the traditional Ratcliff et al. (1990) design, we did not include a pure strong condition composed entirely of strong items. This is because recently it has been discovered that the decrease in performance through the course of recognition memory testing (e.g., Peixotto, 1947) is sensitive to the list composition of the test list, with pure strong lists showing a lesser rate of decline in performance than strong items on mixed lists (Kiliç, Criss, Malmberg, & Shiffrin, 2017). Given that this factor selectively harms mixed strong items, it can artifactually induce an LSE (Mixed strong d' > pure strong d'). In our design, weak items in the mixed lists and pure weak lists are matched on both retention interval and test position. In addition, the strong items are not tested until after the weak items on the mixed list, such that the block of weak items on the mixed list has an identical strength composition to the weak items on the pure weak list. As we will demonstrate later in the text, the models under consideration do not make different qualitative predictions for pure weak vs. mixed weak items and mixed strong vs. pure strong items. Under conditions where an LSE is predicted, the models predict PW d'>MW d' and MS d'> PS d', and likewise when a null LSE is predicted, the models predict PW d' = MW d' and MS d' = PS d'.

We hypothesized item recognition should show a null LSE ($d'_{item,PW} = d'_{item,mixed}$), because the null LSE in item recognition has been replicated quite extensively in the literature, even in cases that have used a strength ratio of 4:1 or greater where strength is manipulated by the number of presentations (Diana & Reder, 2005; Kahana et al., 2005; Ratcliff et al., 1990). Our model predicts a null LSE in both item recognition and source memory when interference among items is low, which is expected to be the case with word stimuli based on prior investigations. Nonetheless, dual process models such as the SAC and Norman and O'Reilly (2003) models predict an LSE in source memory while predicting a null LSE in item recognition. In addition, we will demonstrate later in the article that the basic version of the REM model predicts an LSE in source memory under conditions where it predicts a null LSE in item recognition, although a version with ensemble representations (Criss & Shiffrin, 2005) predicts no LSE in source memory. Thus, there are some models that predict that source memory should be susceptible to an LSE while item recognition should yield a null LSE.

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 across experiments. In Experiment 1, participants were tested on either item recognition or source memory for each studied list, but not both. After completing each study list, they were post-cued on the task to be performed at test. This experiment found a positive LSE in source memory ($d'_{source,PW} < d'_{source,mixed}$) but not in item recognition. Although this result is contrary to the predictions of the Osth and Dennis model, the model of Hautus et al. (2008) provides an alternative explanation in terms of decision processes. In their model, source memory judgments are only elicited for recognized items, while guesses are elicited by unrecognized items because participants do not attempt source retrieval on items they do not recognize. A number of experiments have established that participants adopt higher decision criteria in conditions of higher list strength (e.g., Hirshman, 1995; Starns, White, & Ratcliff, 2010; Stretch & Wixted, 1998); our experiments were no exception, with HR for once presented items and FAR being lower in the mixed list. Under the Hautus et al. (2008) model, a higher decision criterion in the mixed list results in a greater proportion of unrecognized items, and thus more items where source retrieval is not attempted, producing an LSE ($d'_{source,PW} < d'_{source,mixed}$) for reasons that are unrelated to interference or 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 LSE in source memory should be reduced or eliminated. Experiments 2 and 3 directly test this possibility. Experiment 2 was nearly identical to Experiment 1 with the exception that it used a conditionalized source memory procedure. During the test phase, 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. Experiment 3 tested both tasks 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" 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. No LSE in source memory was observed in Experiments 2 and 3.

In each experiment we aimed to collect around 80–90 participants to be consistent with previous sample sizes we have employed in list strength designs (Osth et al., 2014; Osth & Dennis, 2014). This is in part due to the fact that observed LSEs are often quite small in recognition tasks (Buratto & Lamberts, 2008; Osth et al., 2014), and thus we wanted ample sample sizes to have the power to detect such effects if they're present. An exception was Experiment 3, where we collected additional participants beyond our goal, as the conditionalization of source memory data on item confidence results in the omission of a significant proportion of responses. Data from each experiment are posted online at https://osf.io/578xj/.

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. 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. We did not screen for the age or gender of the participants, intact colour vision, or whether participants were native English speakers in this experiment or in subsequent experiments.

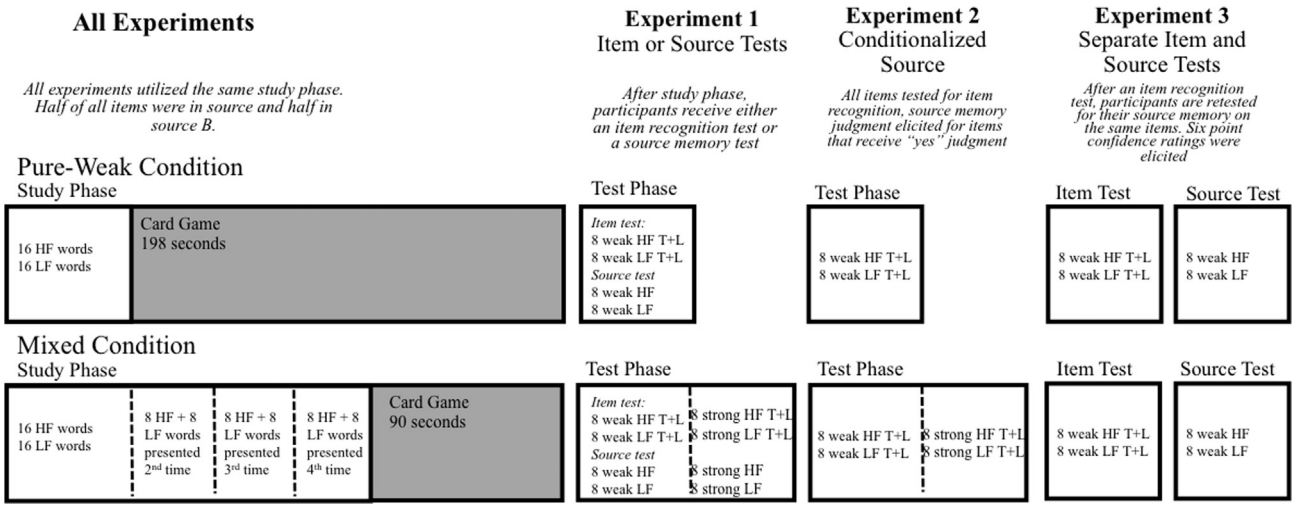

Fig. 2. Diagram of the experimental procedure and how the test phase differs across each experiment. Notes: HF = high frequency, LF = low frequency, T = targets, L = lures.

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 to 9 letters and 1 to 2 syllables in length. All plurals or derivational variants of words were excluded. We attempted to equate for the mean number of neighbors using N-Watch (Davis, 2005). The number of neighbors ranged from 0 to 12 for the high frequency words (M = 1.988) and from 0 to 8 for the low frequency words (M = 0.968). While this was significantly higher for HF words (Z = 4.87, $p < 0.001$), the means were less than one SD apart (SD_{HF} = 2.40, SD_{LF} = 1.42).

Procedure

A diagram of the basic procedure for each experiment can be seen in Fig. 2. During the study phase of both the pure weak and mixed 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 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 198 s in the pure weak condition and 90 s 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. 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 in both tasks to remind the participants of the response keys; in the source memory task these buttons were in the same color as the studied sources to additionally remind participants of the color dimension in their source judgment and were in a similar position as their studied location (e.g., 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 (Peixotto, 1947). 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. Pure weak test lists were composed of 32 trials for item recognition and trials items in length for

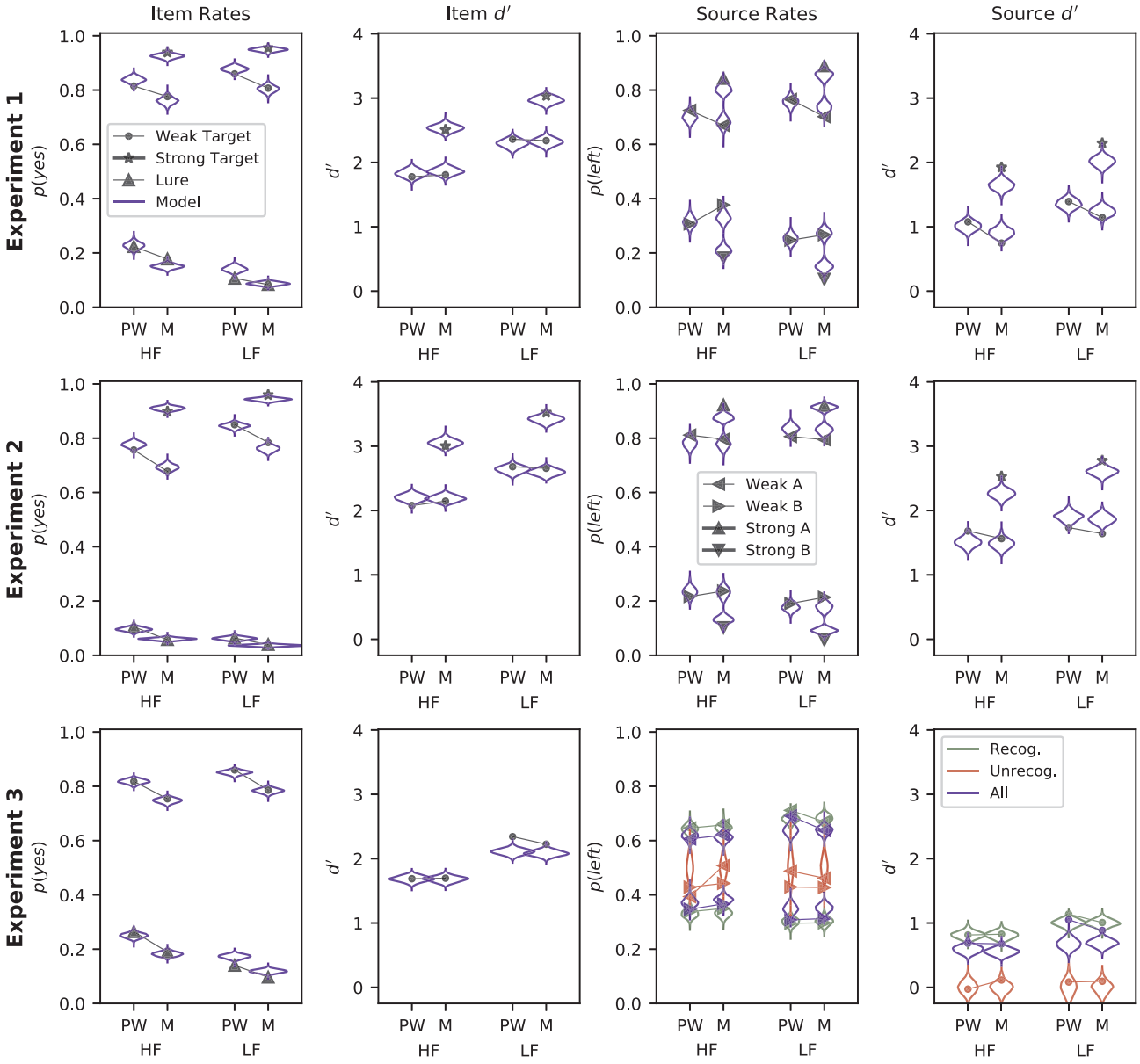

Fig. 3. Group averages along with model predictions for Experiments 1 (top row), 2 (middle row), and 3 (bottom row) for the item recognition task (left two columns) and source memory tasks (right two columns). The posterior predictive distribution of the model is depicted using violin plots. Experiment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

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. None of the words were repeated across the study-test cycles. Half of the study-test cycles were in the pure weak condition while half were in the mixed list condition, resulting in two iterations for each condition in each task.

Results

Results from Experiment 1, along with the other two experiments, can be found in Fig. 3. 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 repeated measures ANOVAs and paired t-tests (on means) 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 0.33–1 range provides only "anecdotal" evidence for or against the null hypothesis, respectively, while BF_{10} in the 3–10 or 0.1–.33 range provides substantial evidence, and $BF_{10} > 10$ or $BF_{10} < 0.1$ provide strong evidence for or against the null hypothesis (Jeffreys, 1961). All statistical tests throughout the article were two-tailed and used default priors.

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 Fig. 3. Source memory analyses were restricted

to d', while for item recognition C, 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 0.5 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 replicated the findings from the list strength paradigm that have been previously reported in the literature. There was no list strength effect, in that d'_{item} did not differ between the pure weak ($M = 2.06$, $SD = 0.84$) and mixed ($M = 2.03$, $SD = 0.77$), $BF_{10} = 0.136$, lists. In accordance with previous results (e.g.; Hirshman, 1995), participants were more conservative with increasing list strength, indicated by a higher criterion C_{item} in the mixed list ($M_{PW} = 0.00$, $M_M = 0.16$, $SD_{PW} = 0.36$, $SD_M = 0.36$), $BF_{10} = 6492.50$. This was reflected in lower HR in the mixed ($M = 0.79$, $SD = 0.14$) than the pure weak ($M = 0.84$, $SD = 0.13$), $BF_{10} = 95.03$ condition along with lower FAR in the mixed ($M = 0.125$, $SD = 0.09$) than the pure weak condition ($M = 0.164$, $SD = 0.12$), $BF_{10} = 8.30$.

We observed a word frequency effect; once presented LF words exhibited higher d'_{item} than HF words ($M_{LF} = 2.32$, $M_{HF} = 1.77$, $SD_{LF} = 0.72$, $SD_{HF} = 0.79$), $BF_{10} = 6.43e + 13$. The locus of the LF advantage was primarily in the FAR, where there were large differences between LF ($M = 0.09$, $SD = 0.08$) and HF ($M = 0.20$, $SD = 0.124$) words, $BF_{10} = 1.94e + 15$, while the HR advantage for LF words was not as strong ($M_{LF} = 0.83$, $M_{HF} = 0.80$, $SD_{LF} = 0.12$, $SD_{HF} = 0.14$), . Item recognition HRs for words presented in each of the two sources was nearly identical ($M_{left} = 0.82$, $M_{right} = 0.81$, $SD_{left} = 0.14$, $SD_{right} = 0.12$), $BF_{10} = 0.095$, which suggests the two sources were represented with equal strength in memory.

In the source memory task, a robust LSE was observed, as reflected in lower d_{source} in the mixed list ($M = 0.94$, $SD = 0.79$) relative to the pure weak list ($M = 1.23$, $SD = 0.76$), $BF_{10} = 213.45$. Post hoc tests revealed that this is especially evident for HF words ($M_{PW} = 1.07$, $M_M = 0.745$, $SD_{PW} = 0.71$, $SD_M = 0.65$), $BF_{10} = 152.46$. While a similar trend was evident for LF words ($M_{PW} = 1.39$, $M_M = 1.14$, $SD_{PW} = 0.77$, $SD_M = 0.87$), the Bayes Factor revealed only very weak evidence in favor of the effect, $BF_{10} = 1.14$. There was also a discriminability advantage for once presented LF words ($M = 1.27$) over HF words ($M = 0.91$), $BF_{10} = 12,440$. The LF advantage also persisted for items that were presented four times ($M_{LF} = 2.30$, $M_{HF} = 1.93$, $SD_{LF} = 0.75$, $SD_{HF} = 1.01$), $BF_{10} = 25.28$.

Discussion

Experiment 1 found a null LSE in item recognition and a robust positive LSE in source memory. These results are inconsistent with the Osth and Dennis (2015) model - although this model can predict an LSE in source memory when interference among the items is high, it similarly predicts an LSE in item recognition under such conditions, which was not found here or in the prior literature. In contrast, these results appear to support the predictions of dual process models such as SAC and the Norman and O'Reilly model. These models predict a positive LSE because recollection is impaired by increasing list strength. In addition, later in the article we will demonstrate that this dissociation is predicted by the basic version of the REM model but not by a later version that employs ensemble representations.

However, there remains an additional decision level explanation of the results of Experiment 1 from the two-dimensional SDT model of Hautus et al. (2008). In this model, source retrieval is only attempted for recognized items; 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 decreased as item confidence decreased, but when items were unrecognized performance abruptly dropped to chance performance (Slotnick & Dodson, 2005). The performance drop was so sharp that this result was unable to be accounted for with other mechanisms, such as correlations between the item and source dimensions. Other investigations have similarly found source memory to be at chance for unrecognized items (Bell, Mieth, & Buchner, 2017; Malejka & Broder, 2016). The psychological interpretation is that participants believe that source retrieval is futile for items they do not recognize, even though they may be able to retrieve some source information for such items.

How could such a mechanism explain the LSE in source memory? In the item recognition data for Experiment 1, HR for once presented items were significantly lower in the mixed list relative to the pure weak list. This pattern has been observed in many list strength paradigms and is often attributed to be the result of participants adopting higher response criteria in conditions of higher list strength (Hirshman, 1995; Starns et al., 2010; Stretch & Wixted, 1998). To understand how a criterion shift could produce an LSE in source memory, consider the HR in the pure weak condition (0.84). Because an average of 16% of items are unrecognized, in the Hautus et al. model we can expect that 16% of the items will receive guess responses on the source judgment and will have 50% accuracy. In the mixed list, in contrast, HRs were lower ($M = 0.79$). This greater proportion of unrecognized items (21%) leads to more guess responses in this condition, producing poorer source memory in the mixed list relative to the pure weak list. Thus, in this model performance would be expected to be poorer in the mixed list even if there are no underlying differences in source discriminability between the pure weak and mixed list conditions. This does not preclude the possibility that there are underlying discriminability differences between the lists *in addition* to performance differences due to the decision mechanism. We investigate this possibility in the next experiment by conditionalizing source memory performance on accurate item recognition.

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, performance should be restricted to cases where participants attempted source retrieval and eliminate guessing based responses. If the positive LSE observed for source memory in Experiment 1 was solely due to this decision mechanism, no LSE should be observed. Experiment 2 affords such a test because 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 from eight in Experiment 1. Because on each study-test cycle

participants engaged in both tasks, this results in three iterations of the pure weak and mixed list conditions for each task (compared to two in Experiment 1).

Results

Results can be found in the middle row of Fig. 3. The item recognition results are consistent with the results of Experiment 1 and prior literature. There was a null LSE on d'_{item} (M_{PW} = 2.38, M_M = 2.34, SD_{PW} = 0.82, SD_M = 0.76), BF_{10} = 0.15. Participants exhibited higher criteria in the mixed list (M_{PW} = 0.24, M_M = 0.48, SD_{PW} = 0.37, SD_M = 0.35), BF_{10} = 6439.47. This was reflected in a lower HR (M_{PW} = 0.803, M_M = 0.731), BF_{10} = 55, 087, and FAR (M_{PW} = 0.084, M_M = 0.047), BF_{10} = 382, 181 in the mixed list condition. LF words were more discriminable than HF words (M_{LF} = 2.64, M_{HF} = 2.08, SD_{LF} = 0.75, SD_{HF} = 0.73), BF_{10} = 3.42e + 18, as LF words exhibited a higher HR than HF words (M_{LF} = 0.82, M_{HF} = 0.72, SD_{LF} = 0.13, SD_{HF} = 0.16), BF_{10} =4.67e + 10, along with lower FAR (M_{LF} = 0.05, M_{HF} = 0.08, SD_{LF} = 0.05, SD_{HF} = 0.08), BF_{10} = 8559. Item recognition HRs were again found to be virtually equal between the two sources (M_{left} = 0.76, M_{right} = 0.77, SD_{left} = 0.15, SD_{right} = 0.13), BF_{10} =0.120.

The source memory results contrasted markedly with those from Experiment 1. There was no effect of list strength on d'_{source} (M_{PW} = 1.79, M_M = 1.68, SD_{PW} = 0.88, SD_M = 0.91), BF_{10} = 0.36. There was no effect of list strength on either HF words (M_{PW} = 1.68, M_M = 1.56, SD_{PW} = 0.97, SD_M = 0.91), BF_{10} = 0.144, or LF words (M_{PW} = 1.73, M_M = 1.64, SD_{PW} = 0.79, SD_M = 0.87), BF_{10} =0.173. Surprisingly, there was no effect of word frequency on d'_{source} for once presented words (M_{LF} = 1.77, M_{HF} = 1.73, SD_{LF} = 0.70, SD_{HF} = 0.83), BF_{10} = 0.183, which is contrary to the results of Experiment 1. However, there was an LF advantage for 4× words (M_{LF} = 2.77, M_{HF} = 2.53, SD_{LF} = 0.75, SD_{HF} = 1.00), BF_{10} = 9.37.

Discussion

We conducted Experiment 2 as a test of whether the positive LSE observed in the source memory task in Experiment 1 was a decision level phenomenon 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 higher incidence of unrecognized items in the mixed list should result in more guessing in that condition. 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 mechanism from the (Hautus et al., 2008) model, we found a null LSE in source memory while finding a null LSE in item recognition. This lack of dissociation between the tasks is consistent with a source memory extension of the Osth and Dennis (2015) model along with a version of the REM model that employs ensemble representations (Criss & Shiffrin, 2005), but is inconsistent with dual process models such as SAC and the Norman and O'Reilly model, which predict a positive LSE in source memory.

We were somewhat surprised to find no effect of word frequency on source memory performance for once presented words despite finding an effect in item recognition. We initially hypothesized that the conditionalized procedure may have eliminated the effect, as the Hautus et al. decision mechanism could similarly produce a source memory advantage for LF words on the basis that they receive less guess responses due to their higher item memorability. However, two pieces of evidence are contrary to that explanation. The first is that an LF advantage was found for 4× words in this experiment. The second is that Glanzer et al. (2004) found an LF advantage in source memory using a similar conditionalized procedure. 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 succession, 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). Because conditionalization can result in the omission of a significant proportion of data, we collected more participants in this experiment than in prior experiments.

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.

Just as in Experiment 2, participants engaged in six study-test cycles (three iterations of each list strength condition).

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 ($d'_{item} < 0.15$); two of these participants had similarly poor performance on the source memory task.

The results can be found in the bottom row of Fig. 3. The item recognition results replicate those of Experiments 1 and 2 and prior results in the literature. There was no LSE on d'_{item} (M_{PW} = 2.01, M_M = 1.96, SD_{PW} = 0.97, SD_M = 0.84), BF_{10} = 0.156. Participants employed higher criteria in the mixed list condition (M_{PW} = −.06, M_M = 0.20, SD_{PW} = 0.35, SD_M = 0.36), BF_{10} = 3.77e + 18.

This was reflected in lower HR ($M_{PW} = 0.839$, $M_M = 0.771$, $SD_{PW} = 0.12$, $SD_M = 0.13$), $BF_{10} = 0.87e + 10$, and FAR ($M_{PW} = 0.165$, $M_M = 0.126$, $SD_{PW} = 0.14$, $SD_M = 0.12$), $BF_{10} = 21,757$, in the mixed condition. Performance was better for LF than HF words ($M_{LF} = 2.28$, $M_{HF} = 1.69$, $SD_{LF} = 0.87$, $SD_{HF} = 0.85$), $BF_{10} = 2.71e + 24$. LF words exhibited higher HR ($M_{LF} = 0.82$, $M_{HF} = 0.79$, $SD_{LF} = 0.13$, $SD_{HF} = 0.13$), $BF_{10} = 160.96$, and lower FAR ($M_{LF} = 12$, $M_{HF} = 0.23$, $SD_{LF} = 0.11$, $SD_{HF} = 0.16$), $BF_{10} = 1.822e + 19$, than HF words. Item recognition HRs were again virtually equivalent between the two sources ($M_{left} = 0.80$, $M_{right} = 0.80$, $SD_{left} = 0.12$, $SD_{right} = 0.12$), $BF_{10} = 0.075$.

For the source memory data, there was weak evidence for a null LSE on d'_{source} ($M_{PW} = 0.86$, $M_M = 0.78$, $SD_{PW} = 0.82$, $SD = 0.77$), $BF_{10} = 0.403$. There was a null LSE for HF words ($M_{PW} = 0.685$, $M_M = 0.676$, $SD_{PW} = 0.74$, $SD_M = 0.73$), $BF_{10} = 0.107$, while LF words showed weak evidence for an LSE ($M_{PW} = 1.05$, $M_M = 0.88$, $SD_{PW} = 0.86$, $SD_M = 0.79$), $BF_{10} = 1.97$. There was a substantial word frequency effect, with LF words exhibiting higher d'_{source} ($M_{LF} = 0.967$, $M_{HF} = 0.681$, $SD_{LF} = 0.74$, $SD_{HF} = 0.83$), $BF_{10} = 328,799$.

We subsequently restricted the source memory data 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 responses. These results are depicted in green in Fig. 3, while the source memory data for unrecognized items are depicted in red. This restriction produced stronger evidence for a null LSE ($M_{PW} = 0.98$, $M_M = 0.91$, $SD_{PW} = 0.87$, $SD_M = 0.86$), as evidenced by a lower BF, $BF_{10} = 0.171$. While HF words still exhibited a null LSE ($M_{PW} = 0.814$, $M_M = 0.826$, $SD_{PW} = 0.83$, $SD_M = 0.85$), $BF_{10} = 0.107$, LF words exhibited weak evidence for a null LSE ($M_{PW} = 1.14$, $M_M = 1.01$, $SD_{PW} = 0.89$, $SD_M = 0.86$), $BF_{10} = 0.423$. Contrary to Experiment 2's results, a word frequency effect persisted after the conditionalization, although the evidence for the effect is weaker than when the analysis is unrestricted ($M_{LF} = 1.14$, $M_{HF} = 0.820$, $SD_{LF} = 0.84$, $SD_{HF} = 0.88$), $BF_{10} = 798.84$.

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 responses. This restriction produced stronger evidence for a null LSE on d'_{source}, $BF_{10} = 0.107$, and extremely similar d'_{source} in the pure weak ($M = 1.04$, $SD = 0.90$) and mixed list ($M = 1.03$, $SD = 0.93$) conditions. Both HF ($M_{PW} = 0.892$, $M_M = 0.951$, $SD_{PW} = 0.88$, $SD_M = 0.91$), $BF_{10} = 0.132$, and LF ($M_{PW} = 1.18$, $M_M = 1.12$, $SD_{PW} = 0.90$, $SD_M = 0.93$), $BF_{10} = 0.130$, words exhibited strong evidence for a null LSE. The word frequency effect again persisted in this analysis, although evidence for the effect is again weaker than in the previous analysis ($M_{LF} = 1.15$, $M_{HF} = 0.922$, $SD_{LF} = 0.90$, $SD_{HF} = 0.92$), $BF_{10} = 182.96$.

Source memory for unrecognized items. Previous investigations have found no source memory for unrecognized items under conditions of unbiased performance (e.g., Malejka & Broder, 2016). While these investigations have typically evaluated whether percentage correct exceeded 0.5, a difficulty with percentage correct is that some participants had more unrecognized items for one source than another, which can produce percentage correct values greater than 0.5 if there is bias toward a particular source. For example, consider if a participant recognized more items that were presented in source A (8 trials) than source B (4 trials), but did not have accurate source memory for any unrecognized items and defaulted to giving a source A response to these items. This would result in 8 trials with an accurate source response (source A items) and 4 trials with an inaccurate source response (source B items), producing a percentage correct of 66% on unrecognized items despite having no source discriminability. While there was no underlying difference in recognition of the source A and source B items at the group level, there were differences in the HR to source A and source B items among the individual participants. An SDT model is robust to this possibility by separating discriminability from bias for each participant.

Unrecognized items (depicted in red in Fig. 3) 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'_{source} 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'_{source}. To partially ameliorate this problem, we collapsed across conditions while calculating d'_{source}. This analysis found that d'_{source} was extremely close to chance for unrecognized items ($M = 0.15$) 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'_{source} that were much closer to zero, while participants with very low numbers of unrecognized items produced d'_{source} 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. These SDT models only made contact with the source memory data; performance on the item recognition task was not modeled to avoid imposing any relationships between the two tasks. 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 uncertainty. 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'_{source} parameter. In the simplest model, d'_{source} 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'_{source} parameter, including a single d'_{source} across all conditions, d'_{source} varying over word frequency conditions ($d'_{source} \sim$ WF), d'_{source} varying over list strength conditions ($d'_{source} \sim$ LS), and d'_{source} varying over all conditions ($d'_{source} \sim$ 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). Relatively non-informative prior distributions were employed for the model parameters; these and other details of the fitting procedure 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.

For the unrecognized items, the results strongly reject the $d'_{source} = 0$ model, indicating source memory for unrecognized items. The preferred

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.

Model	N	Unrecognized		Recognized	
		WAIC	Prob.	WAIC	Prob.
$d'_{source} = 0$	10	4375	0	11,827	0
Single d'_{source}	11	4365	0.003	9714	0
$d'_{source} \sim$ LS	12	4372	0	9795	0
$d'_{source} \sim$ **WF**	**12**	**4353**	**0.9997**	**9653**	**1.0**
$d'_{source} \sim$ LS, WF	14	4370	0.0005	9759	0

The winning model is depicted in boldface.

model for both unrecognized and recognized items shows that d'_{source} only varies across HF and LF words but does not vary across the list strength conditions. The posterior distribution for the group means of $d'_{source,HF}$ and $d'_{source,LF}$ can be seen in Fig. 4. d'_{source} is above chance for both HF ($M = 0.098$, 95% highest density interval, or HDI: $[0.019, 0.17]$) and LF ($M = 0.208$, 95% HDI: $[0.089, 0.325]$) words. These estimates stand in stark contrast to the d'_{source} estimates for recognized items, which are considerably higher for both HF ($M = 0.82$, 95% HDI: $[0.70, 0.95]$) 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.

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 LSE in source memory. This stood in contrast to the results of Experiment 1, which found a positive LSE. However, when the data were conditioned on recognized items and again conditioned on high confidence item responses, stronger evidence for a null LSE 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'_{source} 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 et al. (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).

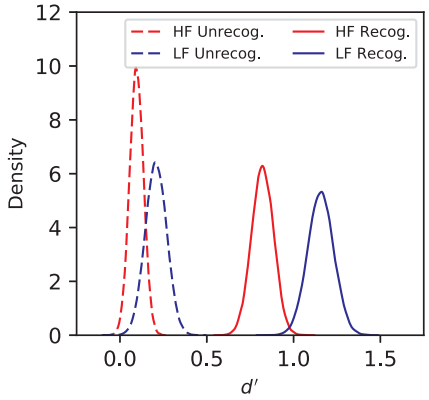

Fig. 4. Posterior distributions of the group means of the d' parameters from the $d' \sim$ WF SDT model applied to the unrecognized items from Experiment 3.

Although the above-chance performance for unrecognized items may appear challenging for the mechanism in the Hautus et al. (2008) model, the paradigm in Experiment 3 may have introduced some confounds. First, presentations of the items in the item recognition test may have increased their memorability and facilitated their recognition on the source memory test, making them less likely to elicit guess responses on the source memory test. The other possibility is that there is criterion variability in the old-new decision (e.g.; Benjamin, Diaz, & Wee, 2009), 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. Given some of these difficulties, the results of this analysis should be interpreted with some caution. In the next section, we evaluate the feasibility of the source guessing mechanism of the Hautus et al. (2008) model within a global matching model of item recognition and source memory.

A global matching model of source memory

Here, 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.

Two factors distinguish source memory from item recognition. First, each source cue (source A and source B) is employed 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 stored associations between items and contexts. The term "context" in episodic memory models is 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 before *in general*; 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, which is defined by the context representation.

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 for this separation 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 have recently found evidence for such three-way bindings using an A-B A-Br paradigm in source memory (Yim, Osth, Sloutsky, & Dennis, 2018). We formalize this conjunctive representation as a mode three tensor, which is a three-way outer product of the I, C, and S vectors:

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

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, item, and source cues into a tensor and using it to probe the memory tensor M:

$$s = (C_s' \otimes I_t' \otimes S_a'). M \quad (2)$$

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 Fig. 5. The similarity 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. Fig. 5 only demonstrates the source A cue; subsequently, the source B cue is applied and the difference between the two memory strengths is calculated. Nonetheless, it is similarly possible to accomplish the task by only using a single source cue (such as the source A cue), where one could respond with the cued source if memory strength exceeded a criterion and respond with the other source otherwise. This seems highly likely when emphasized by the instructions ("Was the word studied in source A?"). Given that our instructions emphasize usage of both source cues, we have adopted that assumption here. In addition, when both source cues are used, there is equal variance between the source A and source B memory strength distributions in the model when each source has equal strength (see Appendix A) which is consistent with observed zROC slopes of one (Glanzer et al., 2004; Slotnick & Dodson, 2005).

Memories fall into one of four categories depending on whether they match or mismatch the item and context cues. Fig. 5 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

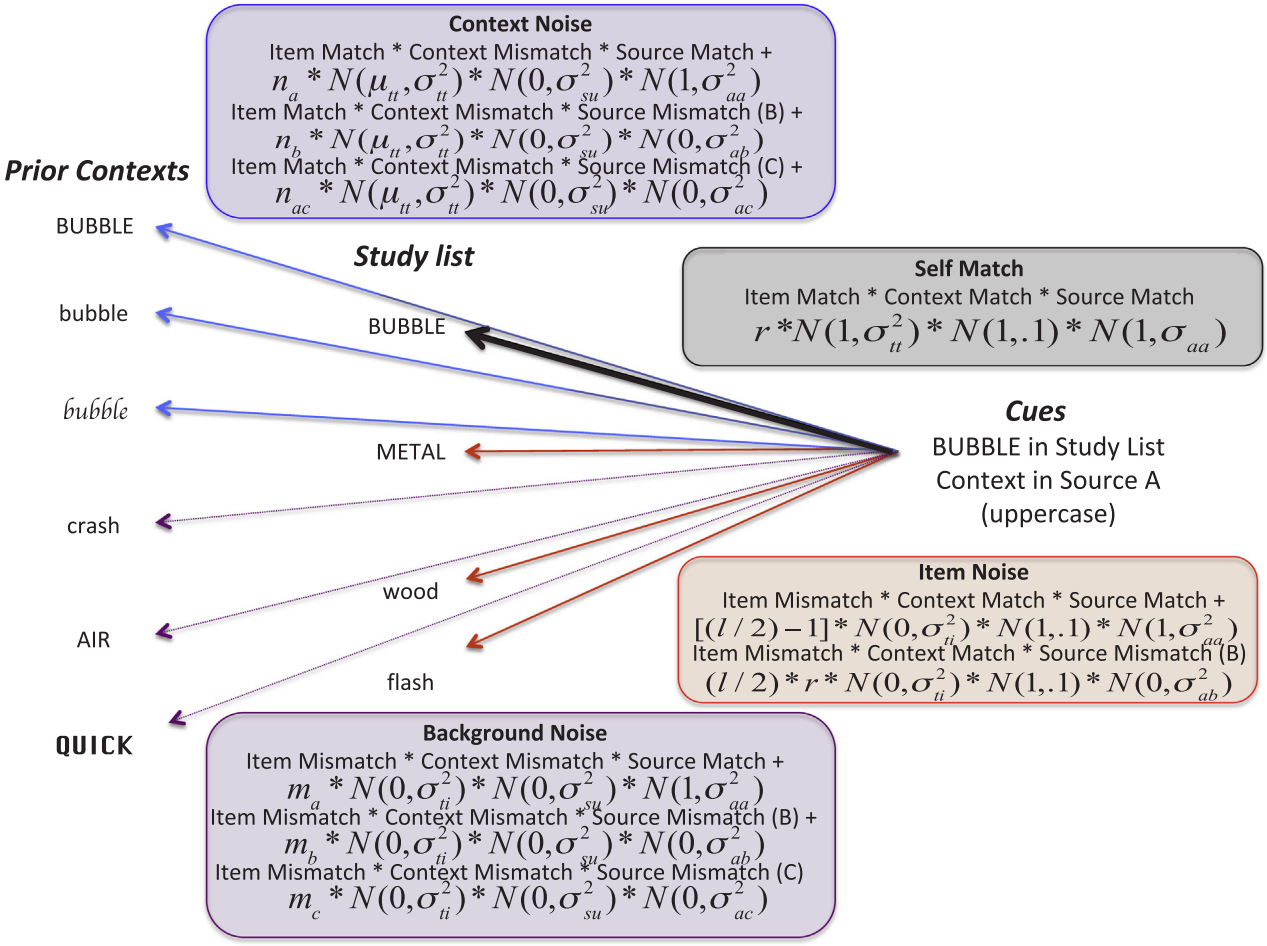

Fig. 5. Illustrative example of the source memory extension of the Osth and Dennis (2015) model. Items in memory are associated with either the study list (right) or prior contexts (left). Items in memory were also studied in source A (denoted by uppercase font), source B (denoted by lowercase font), 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.

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 0.1 and represents the ability to reinstate the study list context. However, in 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, and varies across weak and strong items in our experiments.

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.

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 Fig. 5, and is critically responsible for the predictions about list strength effects. Item noise is scaled by the learning rate r, such that stronger items produce more interference, producing an LSE. 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 interference from other items on the list. Psychologically, this can correspond to the degree to which items are similar to each other; if items are completely dissimilar to each other, there is no interference among the studied items and therefore the number or strength of other items cannot influence performance. Osth and Dennis (2015) found that the list length effect and LSE of varying magnitudes across stimuli reflected different values of σ_{ti}^2, with words exhibiting very low item noise and confusable stimuli such as fractal images exhibiting relatively high values.

The remaining interference sources come from pre-experimental memories. *Context noise* refers to interference from prior occurrences of the item cue. 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 Fig. 5 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 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.

The new parameters of the model correspond to the source matches and mismatches. The match on the source dimension is a draw from a normal distribution with mean 1 and variance σ_{aa}^2, which reflects the noise in the match of a source to its own representation. As this parameter increases, interference increases from memories that were bound to the source A context if source A is used as a cue. 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, such as when source A is used as a cue but is matched to items that were associated with source B. Half the items are expected to match the source cue while half should mismatch. While Fig. 5 illustrates the case where the source associated with the self match matches the source cue (source A), it will mismatch when the source B cue is used.

The tensor model described above applies to source memory; how might predictions be derived for item recognition where only two cues are required (item and context), 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.

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 the 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), 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. Log likelihood ratios are compared to a decision criterion in each task (Φ_{item} and Φ_{source}) to produce a decision.

Table 2
Description of each of the model's parameters, including their boundaries and which conditions they change.

Param	Bounds	Description
r_{weak}	0: ∞	Learning rate for once presented items
rs	1: ∞	Strength factor for strong items; multiplied by r_{weak} to produce learning rate for strong items (r_{strong})
σ_{tt}^2	0: ∞	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: ∞	Number of prior occurrences of HF items in memory. Increases context noise for HF words
σ_{ti}^2	0: ∞	Item mismatch variability: Increases item and background noise in the model
σ_{su}^2	0: ∞	Context mismatch variability: Increases context and background noise in the model
σ_{aa}^2	0: ∞	Source match variability: Increases noise for matches on the source dimension. Note that $\sigma_{bb}^2 = \sigma_{aa}^2$
σ_{ab}^2	0: ∞	Source mismatch variability: Increases noise for mismatches on the source dimension. Note that $\sigma_{ba}^2 = \sigma_{ab}^2$
σ_{ac}^2	0: ∞	Source mismatch variability: Increases noise for mismatches for sources outside of the experiment. Note that $\sigma_{ac}^2 = \sigma_{bc}^2$
Φ_{item}	−∞: ∞	Response criterion for item recognition (0 = unbiased)
Φ_{source}	−∞: ∞	Response criterion for source memory (0 = unbiased)

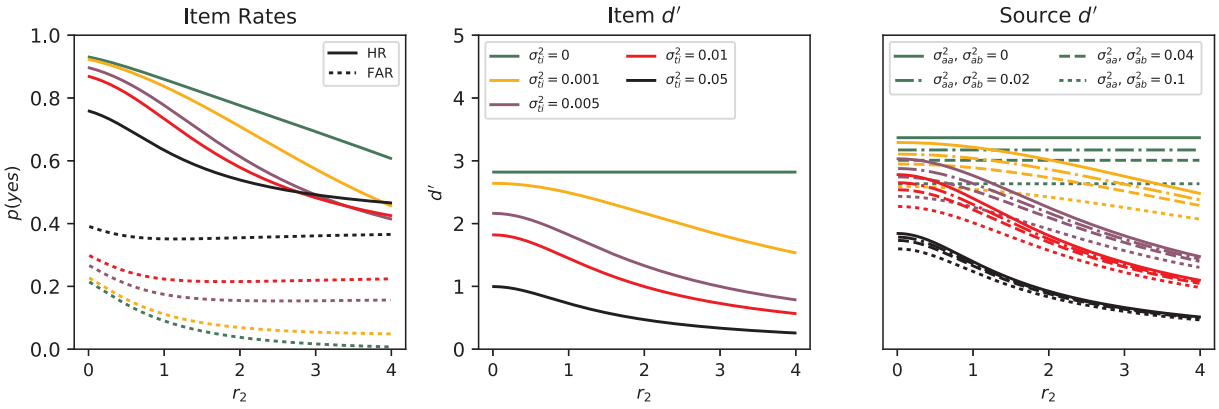

Fig. 6. Model predictions for the list strength paradigm for item recognition (HR/FAR in left panel, d' in right panel), and source memory (d'; right panel). See the main text for more details. Model parameters: $r = 1.0$, $\sigma_{ss}^2 = 0.1$, $\sigma_{tt}^2 = 0.01$, $\sigma_{su}^2 = 0.015$, $\sigma_{ac}^2 = 0.25$, $n = 5$, $\Phi_{item} = 0$, $\Phi_{source} = 0$

Predictions for the list strength paradigm

Predictions for the paradigm can be seen in Fig. 6. Predictions in item recognition (HR and FAR in the left panel, d'_{item} in the middle panel) and source memory (d'_{source}'; right panel) were generated with a range of different values of the item mismatch variability parameter σ_{ti}^2, which governs the total amount of item noise, along with a range of different values of the source match variability parameter σ_{aa}^2 and source mismatch variability parameter σ_{ab}^2, which increase the interference of the source representations. To simplify the predictions, a fixed background noise of 0.05 was assumed for item recognition and 0.1 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 focal tested 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 0.01 and 4.0, to evaluate the extent to which the learning rate of these items interfered with the other half of the list items. One should note that this figure encapsulates predictions for the entire list strength design. When r_2 is less than 1.0 (the learning rate of the focal items), this is analogous to a mixed strong condition because the interfering items are weaker than the focal items. When r_2 is greater than 1.0, this is analogous to a mixed weak condition because the learning rate of the interference items exceeds that of the focal items. When $r_2 = 1$, this is akin to a pure list, because both the focal and interference items have the same strength (this would be observed if all items were studied with the same presentation rate or number of repetitions).

One can see from the figure that when $\sigma_{ti}^2 = 0$ (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 a null LSE is predicted in item recognition and 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_{aa}^2 and σ_{ab}^2) are increased (the dashed and dotted lines show higher values of σ_{aa}^2 and σ_{ab}^2). 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_{ti}^2 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 $\sigma_{ti}^2 > 0$. The parameters that govern interference among the sources (σ_{aa}^2 and σ_{ab}^2) do not appear to strongly interact with the σ_{ti}^2, implying that the list strength predictions are mostly reliant on the σ_{ti}^2 parameter and not on the parameters governing interference from the source representations. This is because each memory's interference is a three way multiplication of the similarity to each cue employed. As the similarity to the item cue approaches zero, the overall similarity should approach zero, and no interference should result.

The model fit

The three experiments from the article provide a rich set of benchmarks for the model, including word frequency, improved performance on strong items, reduced FAR in conditions of higher list strength, ROC shapes (Experiment 3), and the presence and absence of LSEs (positive LSE in Experiment 1 for source memory, null LSEs elsewhere). 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. In addition, model code along with the DE-MCMC software are available online at https://osf.io/578xj/. 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 others are held constant. In addition, parameters corresponding to the source cues (σ_{aa}^2, σ_{ab}^2, and σ_{ac}^2) could be fixed in future applications of the model based on the parameter estimates here.

We fit two implementations of the model that varied with their

assumptions about the data from Experiment 1. The first model employs the aforementioned mechanism from the Hautus et al. model that 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 retrieval from the model and according to a guessing process where all source memory decision probabilities were fixed to 0.5. The former process is weighted by the HR while the latter is weighted by the miss rate, both of which are generated by the model's fit to the item recognition data. 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 from each participant and averaging across participants. Fig. 7 shows the results from each model for Experiment 1. Both models exhibit very similar predictions with the exception of the LSE in source memory. In particular, the non-mixture model fails to predict the LSE; equivalent performance is predicted for the pure weak and mixed list conditions. This is because the item recognition data along with the null LSE 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 the higher retrieval expectations in the mixed list reduce the HR in the mixed list, which consequently produces a greater degree of guessing in the source memory task due to the greater number of unrecognized items. Aside from these differences, the two models yielded extremely similar predictions.

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.

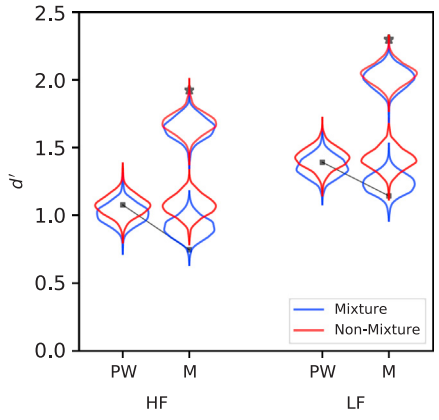

Fig. 7. Group averages along with model predictions for the source memory d' in Experiment 1 for the mixture and non-mixture models.

Aside the source memory LSE in Experiment 1, the model does a very good job of addressing the data from each experiment. The model also produces a null LSE in item recognition (all experiments) along with the null LSE 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. 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 Fig. 3). The model also appears to be providing a strong account of the word frequency effect in both tasks. 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.

The model additionally accounts for the reduced HR and FAR in the mixed list relative without requiring a criterion shift. This follows from the likelihood ratio decision mechanism (Glanzer et al., 2009; Osth et al., 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 Fig. 8 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. Overall, the model is providing a good account of the ROC shapes across both tasks.

To get a sense of why the null LSE 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 Fig. 9. 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 due to minimal interference among word stimuli,

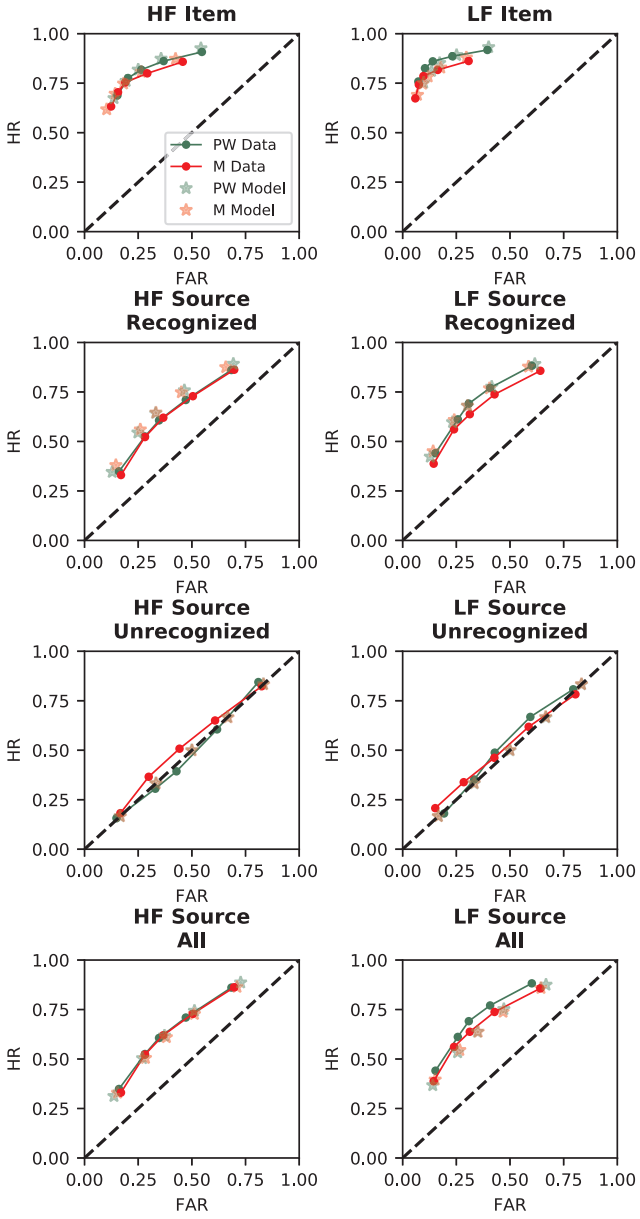

Fig. 8. Group averaged ROC data and model predictions for Experiment 3 for item recognition (top row), along with source memory for recognized items (second row), unrecognized items (third row), and collapsed across recognized and unrecognized items.

similar to prior results (Osth & Dennis, 2015; Osth et al., 2018).

General discussion

The list strength paradigm asks whether strengthening some memories can cause forgetting of other memories. Many experiments have found that this is not the case in item recognition; increasing the list strength of a set of items does not impair discrimination of other memories. However, to our knowledge this question has not been asked in the domain of source memory. Three experiments tested participants on both item recognition and source memory. In Experiment 1, we found strong evidence for the presence of an LSE in source memory while finding a null LSE in item recognition. However, it was unclear as to whether the observed LSE in source memory was due to a decision phenomenon where unrecognized items elicit guess responses (e.g.; Hautus et al., 2008). Specifically, if this mechanism is correct, the lower frequency of recognized items in conditions of higher list strength implies that there should be greater degrees of guessing on source memory decisions in conditions of higher list strength. Thus, in Experiments 2 and 3, we implemented procedures that enabled us to conditionalize source memory performance on recognition on an item-by-item basis, which should reduce the differential degrees of guessing across the two list strength conditions. Both of these experiments observed null LSEs after the conditionalization, suggesting that there is no LSE in underlying source memory discriminability.

Tests for the presence of an LSE in source memory are critical, as they constrain the extension of current models of recognition memory to source memory. In this article, we presented an extension of the Osth and Dennis (2015) global matching model to the case of source memory. This model claims that at study, a conjunctive binding of the item, list context, and source is stored in memory. 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. Simulations in Fig. 6 revealed that the model predicts that both item recognition and source memory should be similarly affected by list strength; a null LSE in item recognition should be accompanied by a null LSE in source memory. The extent to which the tasks are vulnerable to LSEs depends on the interference among the studied items; when this is minimal, a null LSE is predicted in both tasks. Prior investigations with this model have measured very low levels of interference from studied items when the items are word stimuli.

The model was applied to all experiments simultaneously using hierarchical Bayesian estimation and provided a good account of the data. The set of the three experiments provided a rich set of benchmarks for the model, which additionally included word frequency effects and ROC shapes in both tasks. When the model was augmented with a mechanism for which guessing was elicited on unrecognized items, it was capable of accounting for the positive LSE in source memory in

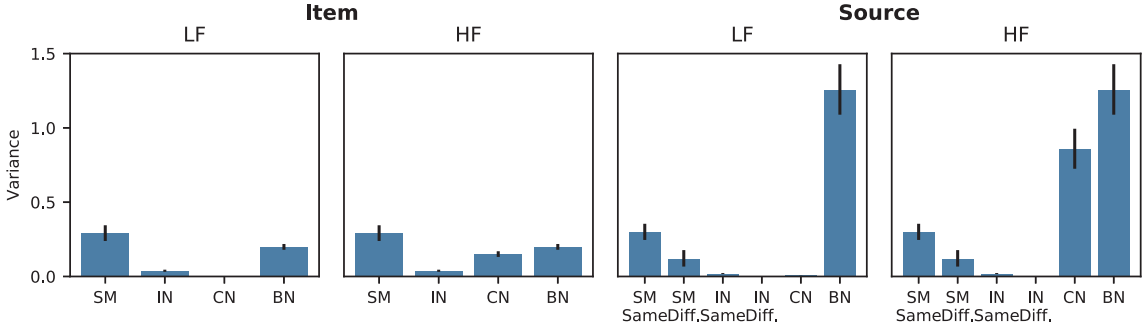

Fig. 9. Memory strength variance from each interference category in item recognition (left two panels) and source memory (right two panels). Note that SM = self match, IN = item noise, CN = context noise, and BN = background noise. In source memory, the self match and item noise components are divided into the contributions from when the same source is used as a cue ("same") and different source is used as a cue ("diff.").

Experiment 1 while accounting for the null LSE in item recognition and the null LSEs in both tasks in Experiments 2 and 3. The model was also capable of addressing the decreased HR and FAR in the mixed list in item recognition, the word frequency effect in both tasks, 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 across conditions. The model accounted for the null LSEs in both tasks because the bulk of the interference contributions appear to come from pre-experimental sources rather than the experimental context itself, similar to previous work, while there was minimal interference among the studied items (Osth & Dennis, 2015; Osth et al., 2018).

Current modeling efforts in the source memory task have largely focused on whether decisions are based on continuous information or discrete states (e.g., Hautus et al., 2008; Klauer & Kellen, 2010). While these models have yielded useful predictions about the relationships between the tasks, they are agnostic with respect to encoding, representation, and retrieval operations, which are often of great interest to memory researchers. The importance of specifying such mechanisms is exacerbated when one considers that dissociations between item recognition and source memory have been heavily investigated in the contexts of aging and amnesic patients (e.g., Addante, Ranganath, Olichney, & Yonelinas, 2012; Ghetti & Angelini, 2008; Janowsky, Shimamura, & Squire, 1989; Johnson, Hashtroudi, & Lindsay, 1993). Application of a mechanistic model such as ours to these populations has the potential to produce more fine-grained conclusions about such differences. For instance, selective deficits in source memory in aging could be due to poorer ability to associate the source features to the items (e.g., Naveh-Benjamin, 2000), greater interference from the source cues, or may indicate a more general memory deficit (e.g., Benjamin, 2010), either due to increased interference or poorer learning. The fact that likelihoods for our model can be expressed analytically make it such that our model can address these questions in a tractable manner, which may be difficult to accomplish in other models that require lengthy simulation times.

While the finding of the null LSE in item recognition and source memory are consistent with our model, they are inconsistent with dual process models such as SAC (Diana & Reder, 2005) and the Norman and O'Reilly (2003) model. Each of these models assumes that recollection should be impaired by list strength. Familiarity-based discrimination is unaffected by list strength, which enables the models to predict a null LSE in item recognition. While the models could perhaps be revised such that familiarity could subserve source discrimination, cued recall additionally shows a null LSE (Wilson & Criss, 2017). In the next section, we discuss how the results compare with REM's predictions which depend on how source information is represented in the model.

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. Criterion variability (e.g.; Benjamin et al., 2009) or learning of the items during the recognition test could have caused recognition during source testing of initially unrecognized items, which would inflate their source memory performance.

REM model predictions for list strength effects in source memory

In the REM model, items are represented as vectors with features sampled from a geometric distribution. During learning, a noisy copy of each studied item is stored in memory as a trace. 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.

Shiffrin and Steyvers (1997) created associations that are concatenations among the to-be-associated items. That is, if an association between "cat" and "dog" is learned and each vector contains 20 elements, these vectors are concatenated to create a vector with 40 elements that corresponds to the association. 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. 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 $c = 0.7$ (noise during encoding), $g = 0.3$ for LF words, and $g = 0.45$ for HF words, while using $g = 0.4$ for source vectors.

Predictions can be seen in the left panel of Fig. 10. We used three different learning rates for the focal (tested) items ($u = 0.12, 0.24, 0.36$) while varying the learning rates of the interfering items across five levels that were lower, equal, or higher than the focal items. Similar to the predictions in Fig. 6, the figure encapsulates the entire list strength paradigm. When the learning rate for interference items is less than the

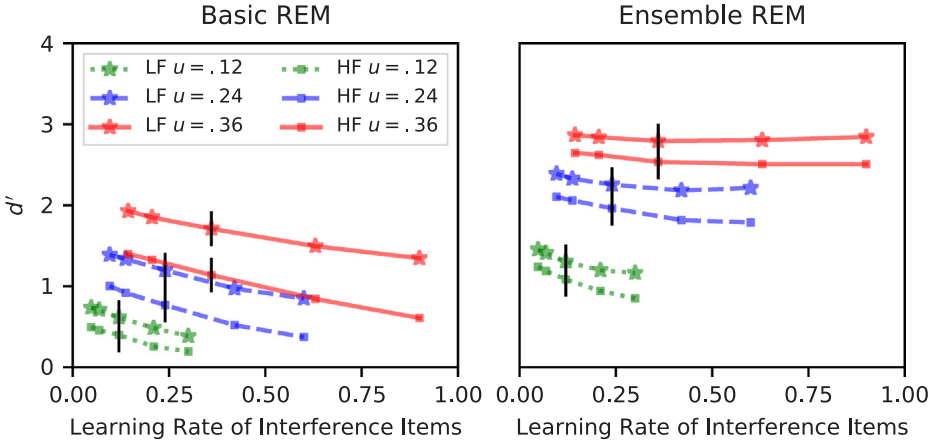

Fig. 10. REM model predictions for our list strength paradigm in source memory. The left panel shows predictions from the simple concatenation model proposed by Shiffrin and Steyvers (1997) while the right panel shows predictions from the Criss and Shiffrin (2005) ensemble features model. The black dashes indicate the point at which the learning rate for the focal items was the same as the learning rate for the interfering items. See the text for details of the simulations.

focal items, this is analogous to a mixed strong condition (interfering items are weaker than the tested items), whereas the case where the learning rate for interference items is greater than the focal items is akin to a mixed weak condition (interfering items are stronger than the tested items). The point where the learning rate for focal and interference items is the same is a pure list (all items have the same strength); these are indicated with black dashes in the figure.

One can see that as the learning rate for interference items is increased, performance on the focal 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 LSE across a range of different parameter values (Criss, 2006; Shiffrin & Steyvers, 1997). This is due to differentiation; as traces in memory are strengthened, their similarity to other items decreases, decreasing item interference as list strength is increased (Criss, 2006). However, Criss (2006) demonstrated that differentiation is most effective when items share few features. 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.

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 $g = 0.4$. 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 Fig. 10. It is immediately evident that the performance of the focal items is considerably flatter as the learning rate of the interfering items is increased over a very wide range, and a null LSE is predicted when the learning rate is sufficiently high ($u = 0.36$).

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 LSE. The functional explanation differs from our model in that REM claims the lack of interference in list strength designs is due to differentiation, while our model claims the lack of interference is due to the bulk of interference coming from memories acquired prior to the experiment.

While the ensemble version of REM is as consistent with our present findings as our own model, other paradigms and manipulations may prove diagnostic for deciding between the models. In item recognition, a core commitment of item noise models is that increasing the similarity among the list items should degrade performance. However, thoroughly controlled manipulations of category length have found that increasing the number of similar items does not decrease discriminability in both two-alternative forced choice (2AFC) testing and when d_A is calculated from confidence ratings (Cho & Neely, 2013). While differentiation predicts decreases in FAR with increasing list strength, REM argues that this is due to the stronger encoding of the traces which will produce a weaker match to lures at test, whereas results have indicated this phenomenon is due to higher retrieval expectations in conditions of higher list strength (Starns et al., 2010; Starns, White, & Ratcliff, 2012; Starns, Ratcliff, & White, 2012). A potentially fruitful avenue of research is evaluating whether these constraints further apply to the source memory task.

In addition, both models clearly diverge with their predictions for manipulations of list length. When there is no interference among the list items, our model predicts no effect of list length on discriminability. In differentiation models such as REM, in contrast, lengthening a study list introduces additional traces which contribute more noise at retrieval (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). To our knowledge, the only investigation which has investigated for the presence of list length effects in source memory is Glanzer et al.’s (2004) Experiment 2, which found large impairments in item recognition and source memory when list length was increased from 50 to 180 items. However, in their design testing immediately followed the study list and study-test lag was matched between the short and long list by testing the final 40 items from each study list. This procedure has been criticized because attentional decreases through the course of list presentation would produce poor encoding of late list items and artifactually induce a list length effect (Dennis et al., 2008; Underwood, 1978). In item recognition, when beginning-of-list items are matched for retention interval instead, either very small (Cary & Reder, 2003) or non-existent list length effects are found (Dennis et al., 2008; Kinnell & Dennis, 2011; Schulman, 1974). Thus, the list length paradigm in source memory merits a re-examination that controls for several of the confounds present in list length designs.

Conclusions

The finding of a null list strength effect in item recognition was very influential, initiating the development of a 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.

Appendix A. Analytic derivation of the source memory model

Following Humphreys, Pike, Bain, and Tehan (1989), we can deconstruct Eq. (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:

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

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

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

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:

$$C_s'\cdot C_s \sim Normal(\mu_{ss}, \sigma_{ss}^2) \quad \text{Context Match}$$

$$C_s'\cdot C_u \sim Normal(0, \sigma_{su}^2) \quad \text{Context Mismatch}$$
$$I_t'\cdot I_t \sim Normal(\mu_{tt}, \sigma_{tt}^2) \quad \text{Item Match}$$
$$I_t'\cdot I_i \sim Normal(0, \sigma_{ti}^2) \quad \text{Item Mismatch}$$
$$S_a'\cdot S_a \sim Normal(\mu_{aa}, \sigma_{aa}^2) \quad \text{Source Match}$$

$$S_a'\cdot S_b \sim Normal(0, \sigma_{ab}^2) \quad \text{Source Mismatch}$$

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 Eq. (5) are substituted into the terms for Eq. (4) to derive mean and variance expressions

for the signal and noise distributions. Because each noise term is 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:

$$E(X_1,...,X_n) = \prod_i E(X_i)$$

$$V(X_1,...,X_n) = \prod_i (var(X_i) + E(X_i)^2) - \prod_i E(X_i)^2$$

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:

$$\mu_{a|a} = r\mu_{tt}\mu_{ss}\mu_{aa} \tag{5}$$

$$\mu_{b|a} = 0 \tag{6}$$

$$\begin{aligned}
\sigma^2_{a|a} = {} & r^2[(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{ss} + \mu^2_{ss})(\sigma^2_{aa} + \mu^2_{aa}) - (\mu^2_{tt}\mu^2_{ss}\mu^2_{aa}) && \text{Self Match} \\
& r^2(l/2-1)[\sigma^2_{ti}(\sigma^2_{ss} + \mu^2_{ss})(\sigma^2_{aa} + \mu^2_{aa})] && \text{Item Noise} \\
& l/2r^2[\sigma^2_{ti}\sigma^2_{ab}(\sigma^2_{ss} + \sigma^2_{ss})] + && \text{Item Noise} \\
& n_a[\sigma^2_{su}(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{aa} + \mu^2_{aa})] + && \text{Context Noise} \\
& n_b[\sigma^2_{su}(\sigma^2_{tt}\sigma^2_{ab})] + && \text{Context Noise} \\
& n_c[\sigma^2_{su}(\sigma^2_{tt}\sigma^2_{ac})] + && \text{Context Noise} \\
& m_a[(\sigma^2_{ti}\sigma^2_{su})(\sigma^2_{aa} + \mu^2_{aa})] + && \text{Background Noise} \\
& m_b(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ab}) + && \text{Background Noise} \\
& m_c(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ac}) && \text{Background Noise}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\sigma^2_{b|a} = {} & r^2[\sigma^2_{ba}(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{ss} + \mu^2_{ss})] + && \text{Self Match} \\
& r^2(l/2-1)[\sigma^2_{ti}\sigma^2_{ba}(\sigma^2_{ss} + \mu^2_{ss})] + && \text{Item Noise} \\
& l/2r^2[\sigma^2_{ti}\sigma^2_{bb}(\sigma^2_{ss} + \mu^2_{ss})] + && \text{Item Noise} \\
& n_a[\sigma^2_{su}\sigma^2_{ba}(\sigma^2_{tt} + \mu^2_{tt})] + && \text{Context Noise} \\
& n_b[\sigma^2_{su}(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{bb} + \mu^2_{bb})] + && \text{Context Noise} \\
& n_c]\sigma^2_{su}\sigma^2_{bc}(\sigma^2_{tt} + \mu^2_{tt})] + && \text{Context Noise} \\
& m_a(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{ba}) + && \text{Background Noise} \\
& m_b[\sigma^2_{ti}\sigma^2_{su}(\sigma^2_{bb} + \mu^2_{bb})] + && \text{Background Noise} \\
& m_c(\sigma^2_{ti}\sigma^2_{su}\sigma^2_{bc}) && \text{Background Noise}
\end{aligned} \tag{8}$$

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

$$\mu_{y-z} = \mu_y - \mu_z$$

$$\sigma^2_{y-z} = \sigma^2_y + \sigma^2_z - cov_{y-z}$$

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:

$$\sigma^2_A = \sigma^2_{a|a} + \sigma^2_{a|b} \tag{9}$$

$$\sigma^2_B = \sigma^2_{b|a} + \sigma^2_{b|b} \tag{10}$$

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:

$$\mu_{old} = r\mu_{tt}\mu_{ss} \tag{11}$$

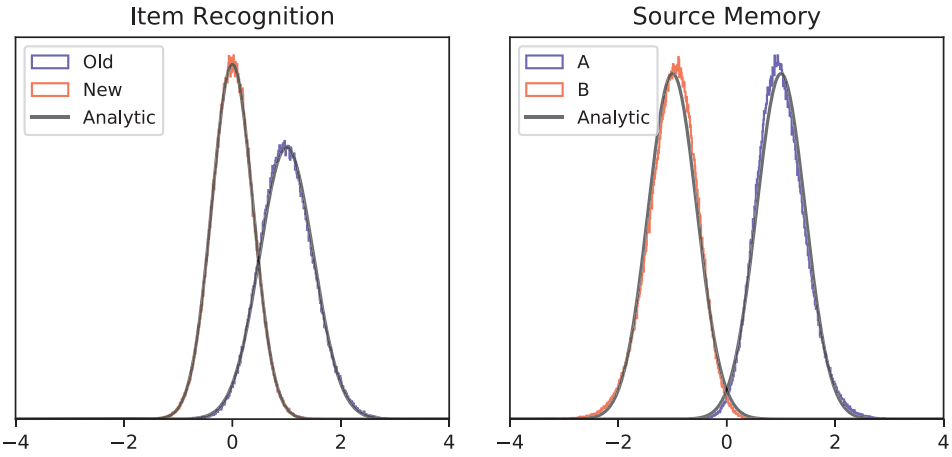

Fig. A.1. Histograms of simulation predictions along with analytic approximations (lines) for the item recognition (left) and source memory (right) model. Model parameters were r, μ_{tt}, μ_{ss}, $\mu_{aa} = 1$, σ^2_{tt}, $\sigma^2_{ss} = 0.05$, σ^2_{aa}, $\sigma^2_{ab} = 0.01$, $\sigma^2_{ti} = 0.001$, n_{aa}, $n_{ab} = 2$, $n_{ac} = 6$, m_{aa}, $m_{ab} = 10$, $m_{ac} = 480$

$$\mu_{new} = 0 \tag{12}$$

$$\begin{aligned}\sigma^2_{old} = {} & r^2[(\sigma^2_{tt} + \mu^2_{tt})(\sigma^2_{ss} + \mu^2_{ss}) - (\mu^2_{tt}\mu^2_{ss})] && \text{Self Match} \\ & r^2(l-1)[\sigma^2_{ti}(\sigma^2_{ss} + \mu^2_{ss})] && \text{Item Noise} \\ & n_{item}[\sigma^2_{su}(\sigma^2_{tt} + \mu^2_{tt})] + && \text{Context Noise} \\ & m_{item}(\sigma^2_{ti}\sigma^2_{su}) && \text{Background Noise}\end{aligned} \tag{13}$$

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

Fig. A.1 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 Eq. (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 et al. (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}:

$$d_{item} = \mu_{old}/\sigma_{new} \tag{15}$$

$$S_{item} = \sigma_{old}/\sigma_{new} \tag{16}$$

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

$$d_{source} = (\mu_A - \mu_B)/\sigma_A \tag{17}$$

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:

$$\mu_{\lambda new} = -\left(\left(\frac{d^2}{2}\right)\left(\frac{S^2+3}{4S^2}\right) + log(S)\right) \tag{18}$$

$$\mu_{\lambda old} = d^2\frac{S^2+1}{2S^2} + \mu_{\lambda new} \tag{19}$$

$$\sigma_{\lambda new} = d\frac{S^2+1}{2S^2} \tag{20}$$

$$\sigma_{\lambda old} = S\sigma_{\lambda L} \tag{21}$$

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

$$\mu_{\lambda A} = d^2/2 \tag{22}$$

$$\mu_{\lambda B} = -d^2/2 \tag{23}$$

$$\sigma_{\lambda A,B} = d^2 \tag{24}$$

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:

$$\mu_{\lambda old} = dd^* \frac{S^2 + 1}{2S^2} + \mu_{\lambda L} \tag{25}$$

$$\sigma_{\lambda old} = S^* \sigma_{\lambda L} \tag{26}$$

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

$$\mu_{\lambda A} = d * d/2 \tag{27}$$

$$\mu_{\lambda B} = -d * d/2 \tag{28}$$

$$\sigma_{\lambda A,B} = d \tag{29}$$

Appendix B. Details of the hierarchical Bayesian fitting procedure

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^2_{ti}) and context (σ^2_{ss}) 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^2_{ss} to 0.1, 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_{ti}), 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.; $n_{LF,aa} = 0.05 n_{LF,item}$), 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:

$$r_{strong,i} = r_{weak,i} * (1 + rs) \tag{30}$$

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. Unlike the r_{weak} parameters, rs 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:

$$d' \sim Normal(M_{d}, \varsigma_{d})$$
$$\phi_{k,j} \sim Normal(M_{\phi kj}, \varsigma_{\phi kj})$$
$$\phi_{k,3,3} \sim Normal(M_{\phi t,3,3}, \varsigma_{\phi t,3,3})$$

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:

$$\phi_{k,1} = \phi_{k,3} - c_{k,1}$$
$$\phi_{k,2} = \phi_{k,3} - c_{k,2}$$
$$\phi_{k,4} = \phi_{k,3} + c_{k,4}$$
$$\phi_{k,5} = \phi_{k,3} + c_{k,5}$$

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:

$$c_{k,1} \sim TN(M_{citem1}, \varsigma_{ck1}, 0, \infty)$$
$$c_{k,2} \sim TN(M_{citem2}, \varsigma_{ck2}, 0, \infty)$$
$$c_{k,4} \sim TN(M_{citem4}, \varsigma_{ck4}, 0, \infty)$$
$$c_{k,5} \sim TN(M_{citem5}, \varsigma_{ck5}, 0, \infty) \quad (31)$$

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

$$M_{\sigma ti,\sigma tt,\sigma su,\sigma aa,\sigma ab,\sigma ac,nHFitem,rweaki,d} \sim Normal(0, 10)$$
$$M_{\phi ti} \sim Normal(0, 1)$$
$$M_{ckj} \sim TN(0.5, 0.5, 0, \infty)$$
$$\varsigma_{\sigma ti,\sigma su,\sigma aa,\sigma ab,\sigma ac,nHFitem,rweakj,\phi ti,ckj,d} \sim Gamma(1, 3)$$

We adopted somewhat stricter priors for the rs and σ_{it}^2 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_{it}^2. To improve estimation of these parameters, we used stricter prior distributions on ς, which place much higher likelihoods on lower values of ς:

$$M_{rs} \sim Normal(0, 0.5)$$
$$\varsigma_{rs,\sigma tt} \sim Gamma(1, 25) \quad (32)$$

For each SDT model, the number of chains was set equal to three times the number of parameters. Sampling began after 6000 burn-in iterations were discarded. The chains were thinned such that only one in every 20 samples was collected; this process continued until 2000 samples were collected in each chain. For the extension of the Osth and Dennis (2015) model to source memory, 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 1500 samples per chain were collected.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jml.2018.08.002.

References

Addante, R. J., Ranganath, C., Olichney, J., & Yonelinas, A. P. (2012). Neurophysiological evidence for a recollection impairment in amnesia patients that leaves familiarity intact. *Neuropsychologia, 50*, 3004–3014.

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. (2010). Representational explanations of process dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review, 117*, 1055–1079.

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.

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

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

Clark, S. E., & 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., & 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.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition, 33*, 1289–1302.

Ghetti, S., & Angelini, L. (2008). The development of recollection and familiarity in childhood and adolescence: Evidence from the dual-process signal detection model. *Child Development, 79*, 339–358.

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

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.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science, 6*(3), 253–271.

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.
Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Source memory impairment in patients with frontal lobe lesions. *Neuropsychologia, 27*, 1043–1056.
Jeffreys, H. (1961). *Theory of probability.* Oxford University Press.
Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28.
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.
Kinnell, A., & Dennis, S. (2011). The li st length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*, 348–363.
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.
Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*, 1–7.
Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*(1), 155–189.
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., & 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.
Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2000*(26), 1170–1187.
Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review, 110*(4), 611–646.
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., & 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.
Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology, 104*, 106–142.
Peixotto, H. E. (1947). Proactive inhibition 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., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535.
Reder, L. M., Nh ouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320.
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.
Schulman, A. L. (1974). The declining course of recognition memory. *Memory and Cognition, 2*, 14–18.
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., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1137–1151.
Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63*, 18–34.
Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40*(8), 1189–1199.
Stretch, V., & Wixted, J. T. (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.
Underwood, B. J. (1978). Recognition memory as a function of the length of study list. *Bulletin of the Psychonomic Society, 12*, 89–91.
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.
Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language, 95*, 78–88.
Yim, H., Osth, A. F., Sloutsky, V. M., & Dennis, S. (2018). Evidence for the use of three-way binding structures in associative and source recognition. *Journal of Memory and Language, 100*, 89–97.
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 characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800–832.