Simulating False Recall as an Integration of Semantic Search and Recognition

Brendan T. Johns (johns4@indiana.edu)
Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth St., Bloomington, In 47405 USA

Abstract

We present a computational model of false recall phenomena. The model is based on an integrated architecture including modules that have been successful at accounting for other types of memory tasks. Words presented in a list are represented by semantic vectors from a co-occurrence model, and are encoded into a composite store. At test, the model generates a list of candidate words from the lexicon and decides which of these words to recall using a recognition process. We show that the model is able to account for a wide range of effects in false recall, including levels of false recall in DRM studies, item-level effects, number of associates, and categorical vs. associative list structure.

Keywords: False recall; co-occurrence representations; memory models; free recall; generate-recognition models

Introduction

The Deese/Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995) has provided fundamental evidence about how humans can remember events that were not stored. In this paradigm, a subject is typically given a list of items to encode. The list contains clusters of items related to a critical lure that is not presented on the list; during subsequent recognition or recall, the critical lure is remembered at similar levels as the encoded items. For example, given *pillow, snore, bed, tired*, etc. to encode, subjects are likely to falsely recall or recognize *sleep*. Exactly what type of information and overlap is necessary between the targets and the critical lure still remains the topic of considerable debate. The DRM paradigm has received much recent attention as a task to understand false memory in applied fields, such as eyewitness testimony.

From a cognitive modeling perspective, the DRM paradigm is particularly challenging because a full understanding of the illusion requires an account of both the *structural* organization of semantic memory and the *process* of memory retrieval. As Estes (1975) originally noted at the outset of cognitive modeling, one cannot study structure or function independently; observed human behavior is an interaction of the two, and to fully understand a cognitive operation, we need a model that explains both structure and function, and how the two interact to produce behavior.

Accounts of false recall have tended to focus on general verbal conceptual frameworks. Recently, the first computational model of false recall was presented (the fSAM model of Kimball, Smith, & Kahana, 2007). While the fSAM model is an excellent step towards a formal framework for understanding false recall, the model focuses on an account of process rather than structure. The model represents the associative connections between words in memory by using association norms, or by hand-fitting representations. While the process of the fSAM model may be a correct one, an explanation of a semantic behavior such as false recall also requires an account of the representation that the process operates upon. We believe that integrating models that create a realistic structural representation with such process models can yield great benefits. If we do not have the correct account of structure when we build the process model, we may have to posit a more complex processing mechanism than humans actually use in order to produce the complex behavior seen in humans. In reality, much of the requisite complexity for a behavior may be coded in the structure of the representation, and a much simpler processing mechanism will suffice to produce the behavior (cf. Jones & Mewhort, 2007).

A promising class of models that can be used to explain the structure of semantic memory are co-occurrence learning models. Examples include Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), the Topics model (Griffths, Steyvers, & Tenenbaum, 2007), and BEAGLE (Jones & Mewhort, 2007). These models build semantic representations for words by observing co-occurrence statistics in a large text base. They have seen considerable success at accounting for a variety of semantic behaviors directly from their representations, such as synonymy test performance, semantic similarity ratings, association norms, and identification times (Johns & Jones, 2008). Due to their success, it seems natural to use co-occurrence representations as structural representations in a process model of false recall.

However, co-occurrence models only provide an account of the structure of memory. To produce sophisticated behavior as in false recall, we require a suitable processing model to interface with this structure. In this paper, we present an account of the process of false recall that operates on a realistic semantic representation learned by a co-occurrence model. We build semantic representations for words using the recent Semantic Distinctiveness Memory (SDM) model (Johns & Jones, 2008). Words presented on a DRM list are retrieved from the SDM mental lexicon and are stored in a composite memory store. At test, the model retrieves a list of candidate words from the lexicon that are similar enough to the composite store, and this candidate list is then used by a recognition module which decides whether or not to recall the word. Our account is an integrated

architecture fusing together models of semantic representation, recognition, and memory search that have proven successful at accounting for other types of memory data. It is the integration of these components that produces our free recall behavior. The integrated architecture of these models is displayed in Figure 1. By integrating different models it is possible to explain more data than any single model can explain by itself. The different aspects of the false recall model will now be described in turn.

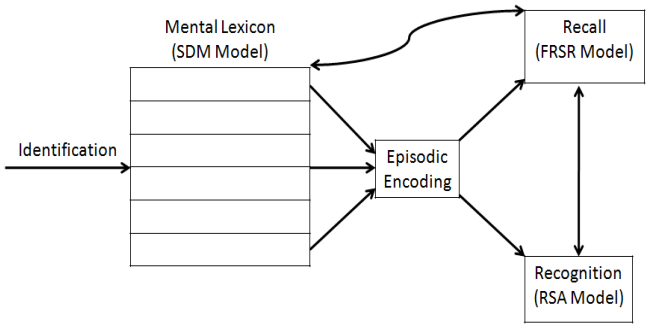

Figure 1. The architecture of the full memory system.

False Recall through Search and Recognition

Our false-recall model is based on classic generate-recognition models (e.g. Kintsch, 1970), in which items are internally generated and are then tested with a recognition check. These simple models will be extended with a mechanism that searches through a fully quantified mental lexicon built by a co-occurrence learning model. Once a search set is created, these words are then tested with a recognition model that has been shown to be susceptible to false recognition. Hence, there are four different modules of the model that we will describe: 1) the lexical representations, 2) the encoding process, 3) the search process, and 4) the recognition process.

1. Lexical Representations from SDM

To simulate the contents of semantic memory, we use representations constructed by the SDM model (Johns & Jones, 2008), a recent co-occurrence learning model. The SDM model produces semantic representations similar to those created by other models such as LSA, but it also simulates the effect of semantic distinctiveness on a word's strength in memory. We have demonstrated, using both a corpus analysis (Johns & Jones, 2008) and an artificial language learning experiment (Recchia, Johns, & Jones, 2008), that words that occur in more semantically distinct contexts are more strongly represented within memory. Johns & Jones (2008) showed that this SDM model produces better fits to both lexical decision and naming times, and produces superior semantic organization compared to other models. In addition, the model can account for semantic isolation effects, semantic similarity ratings, and word association norms.

As in other co-occurrence learning models, SDM builds a term-by-document matrix from a text corpus. The modification that the SDM model makes, in comparison to other co-occurrence models, is the type of information added into the word-by-document matrix: instead of raw frequency, it uses a semantic distinctiveness (SD) value representing how distinct the current context is compared with the previous contexts that the word has occurred in. The first step in computing this SD value is to create a 'context' or 'document' vector, which we call a composite context vector (CCV). This vector represents the meaning of the current context. For each word that occurs in a document (W_1,...,W_N), the word's vector is added into the composite vector. Formally, this is:

$$CCV = \sum_{i=1}^{N} T_i \quad (1)$$

where N is the set of words in the document, and T_i is the memory trace corresponding to word *i*. The next step is to compute a similarity value (given by a vector cosine) between each word that occurs in the context and the CCV. This similarity value is then transferred through an exponential probability density function to give an SD value:

$$SD = e^{-\lambda * cosine(word, CCV)} \quad (2)$$

where λ is a fixed parameter with a small positive value; as λ is increased the difference in the value of high vs. low similarity contexts is accentuated. This SD metric is the value added into the memory slot for that word and context. A context with a high SD value means that it is more distinct compared with the other contexts that a word has appeared in. This practice gives greater salience to more unique contexts, in terms of the word's magnitude, than redundant contexts.

These SDM representations will be used as the lexical structure that drives recall. SDM representations are sparse vectors, in which non-zero values contain a number between 0 and 1 that represents how important that particular context was to forming the semantic representation for that word. Even though the typical practice is to reduce the dimensionality of these vectors using some type of vector reduction technique, as is done in LSA (Landauer & Dumais, 1997) and the Topics model (Griffiths, et al., 2007), we use the raw episodic traces. In a test of the raw vectors against LSA, we find that the model attains a better fit to semantic similarity ratings and backward association strength, and the model only does slightly worse on forward association strength. In the following simulations, the SDM lexicon was built by training on the TASA corpus, and the vectors will have a resulting dimensionality of 36,700.

2. Encoding Process

We use a single composite vector to represent a study list. The representations for each word seen in a study list are retrieved from the SDM mental lexicon, and are summed into a composite vector, which represents of the "gist" of all words on the list. Word vectors are first normalized so that each word adds approximately the same amount of

information, and each is weighted by a uniform random number between 0 and 1 prior to addition to simulate encoding failure.

The use of a composite vector is one of the reasons that this model will be able to account for a variety of false memory results. It is based on the assumption that humans encode the meaning of items in the context of the other items encoded. From this perspective, the task of recall may involve a process that determines if a particular word's semantic representation is coherent with the gist representation that was seen in the study list. This approach has some similarities with the claims of Fuzzy Trace Theory (FTT; Brainerd & Reyna, 2002), which proposes that (among other things) memory stores 'gist' traces of events, and these are traces that capture the meaning of an episode, without specific perceptual features. In a similar way, our model encodes a composite vector of all the words that occur in a specific study list, and this vector represents the gist of the study list.

3. Search Process

To internally generate words to test with recognition, a searching mechanism within the mental lexicon is employed. The SDM lexicon contains approximately 70,000 words. Hence, the searching process is a difficult one: it requires extraction of the words that occurred on the list (and are encoded within the composite vector), whilst ignoring all other words. Due to the immensity of this searching task, the word representation must be of sufficient resolution, and the SDM model contains this necessary structure.

Firstly, we define the similarity between a word in the lexicon and the study list's composite trace as the cosine between their respective vectors. This value is then converted to one minus the magnitude of the word under consideration divided by the maximum magnitude in the lexicon (approximately 1000). Formally, this is

$$Sim = \cos(word, probe) * \left\lfloor 1 - \left(\frac{len(word)}{\max} \right) \right\rfloor, \qquad (3)$$

where *len* returns the magnitude of the word in memory.

This similarity value is then used to drive the searching process. This process is simply based on classic signal detection: If the similarity between a word in the lexicon and the composite is greater than a criterion, then the word is added into the search set. This criterion is fixed at 0.1 across all our simulations. The criterion seems intuitively low because the SDM vectors are sparse vectors, hence, the cosines that are taken with this model tend to be low.

4. Using RSA for the Recognition Process

Once the search set is compiled, we need a decision mechanism to determine if the retrieved words actually occurred. This is necessary because the model does not store any item-level information, so it is not possible to conduct item-to-item comparisons. Instead, a process is necessary to determine if the retrieved word's semantic representation is coherent with the study list's representation.

In Johns & Jones (2009) we describe a recognition model that is designed to do exactly this. This model also uses the semantic representation that the SDM model creates and encodes a study list in the same manner as described above. There are two main aspects to this recognition model, *amplification* and *decision*.

a) Amplification The recognition model is based on an analogy to amplification. For each item in the search set, the recognition attempts to amplify the word's representation in the composite memory representation. This is accomplished in two ways - by adding probe information into the composite and by removing contradictory information. How efficiently a candidate word's semantic information is amplified within the composite may be viewed as confirmatory information (signal), and how much mismatching information is amplified (noise) may be viewed as contradictory information. Contradictory information is taken out by simply multiplying the memory vector by a uniform random number between 0 and 1 at each location where the probe word contains no information (i.e. where the probe vector is 0). This process causes the probe word's representation to increase in the memory store. This process is intimately tied to the decision process described below because how efficiently the word is amplified within memory determines the decision that is made.

b) Decision Two different types of information are used by the RSA model to decide whether the candidate word was encoded or not: 1) similarity information and 2) contradictory information. Similarity information is simply assessed with a cosine between the probe vector and the composite vector. If this cosine exceeds a certain criterion (set at 0.991 in the following simulations) then an 'old' decision is made.

Contradictory information is the amount of information the model has that the word did not occur in the study list. This information is used to make 'new' decisions, and is done by taking the absolute difference between the defining portions of the probe and the corresponding locations within the memory vector, and dividing this summation by the magnitude of the probe. The resulting value is between 0 and 1, where it will be 0 if all of the probe information is contained in memory, and it will be 1 if none of the probe information is contained within memory. Because contradictory information decreases across iterations, this value is a running count. If this count exceeds a criterion (set at 3.9 in the following simulations), then a 'new' decision is made. A detailed formal treatment of both the amplification and decision process can be found in Johns & Jones (2009).

Discussion

This model is based off of classic generate-to-recognition models (e.g. Kintsch, 1970) and utilizes a mental lexicon built by a co-occurrence learning model. At study, the presented list is encoded into a composite memory vector. A searching process then utilizes this composite vector to search through the mental lexicon and pull out the most similar words to the composite, in order to create a search set. Once this search set has been created, the words in this set are given to a recognition model, which decides whether or not to recall a word.

The parameter space for this processing model is very simple – there are only three fixed parameters. These parameters are not manipulated across the different simulations, so there is very little complexity actually built into the processing model. Instead the emphasis in this model is based on the contents of memory for different experiments, and not processing differences across tasks.

Simulations

The methodology that we use in simulating false recall results is very simple: we take the words used in a specific experiment, retrieve the vector representations from the SDM model, encode the words in a study list to a composite representation, and feed this composite to the search and recognition processes.

Simulation #1: Levels of False Recall

Different levels of false recall are observed in different DRM lists. Here we simulate three different sets of DRM lists: 1) the DRM lists from Roediger & McDermott's (1995) classic study, 2) the extended DRM list set from Stadler, Roediger, & McDermott (1999), and 3) the more variable lists from Gallo & Roediger (2002). The levels of veridical recall were also tested to ensure that the model is attaining true recall levels as well as false recall levels.

Method The DRM lists for the above described studies were attained from the specified papers. One list (that for *man*) was excluded because it was in the stop list that the SDM model was trained with. For a single trial, four DRM lists were randomly selected and added into the composite. Then the proportion of studied items recalled was recorded, as well as the number of critical lures falsely recalled on each trial. In total, 250 trials were run for each set of DRM lists. This limited number of trials was conducted because of the large amount of computation that this model requires.

Results The levels of false and veridical recall across the different DRM lists are displayed in Figure 2. This figure shows that the model attains a very good approximation to the levels of recall across the different list sets. Also, the level of non-critical word intrusions across the different list sets was also recorded. For the lists from Roediger & McDermott (1995) 2.2 non-critical words intruded on average. In the Stadler, et al. (1999) lists, 1.8 words intruded, and in the Gallo & Roediger (2002) 3.5 words intruded. These predictions are slightly high compared to the empirical studies, but considering the massive search task that the model must undertake (searching through 70,000 words), the pattern is nonetheless impressive.

However, this comparison only provides qualitative evidence that the model is attaining false recall levels equivalent to those observed in experimental data. As Stadler, et al. (1999) and Gallo & Roediger (2002) have shown, there is considerable variability in the amount of false recall across lists within an experiment. Because our model possesses individual representations for each word, it is possible to measure the different levels of false recall that are seen for particular critical words in the model and compare these quantitatively with empirical results.

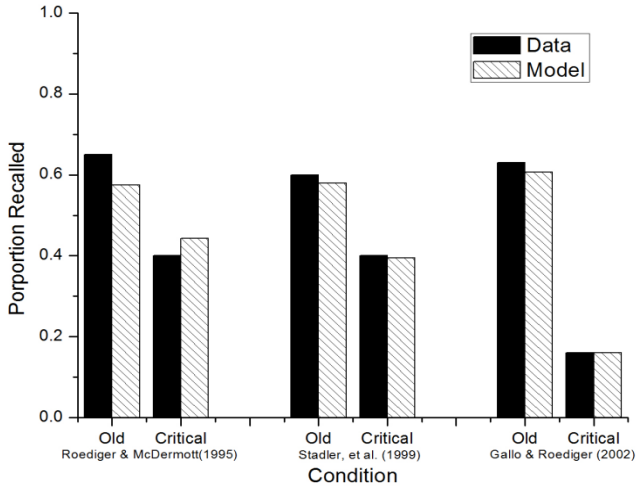

Figure 2. The simulated levels of veridical and false recall and the corresponding empirical results.

Simulation #2: Item-Level Analysis

Stadler, et al. (1999) and Gallo & Roediger (2002) have both published the levels of false recall observed with different DRM lists. As both of these studies show, there is considerable variability in the levels of false recall elicited by different DRM lists. To test the model's quantitative predictions, we correlated the levels of false recall for the model and data using the same words from each experiment.

Method 54 lists from Stadler, et al. (1999) and Gallo & Roediger (2002) were obtained from these studies. Again, four DRM lists were added into a single composite vector and the levels of false recall for the different critical words were attained. In total, 250 trials were simulated for both the lists from the Stadler, et al. and the Gallo & Roediger study.

Results Across the 55 lists (with repeats removed), a significant correlation of $r = 0.496$, $p < 0.001$ was obtained between the model's predictions and the behavioral data. Hence, it appears that the model is producing relative levels of false recall across different critical items that shows strong correspondence to the false recall levels in humans.

If the five lists that the model does worst on (*king*, *rough*, *needle*, *smell*, and *health*) are removed, then the correlation increases to an $r = 0.675$, $p < 0.001$. There is no principled reason to remove these items, but it does show that for the majority of lists the model is giving a good approximation.

However, this simulation does not rule out the possibility that these critical words are being recalled not due to semantic similarity, but to some other factor in their representation (e.g. frequency). To further demonstrate that the amount of semantic information about a specific word contained in memory is driving false recall, we conducted a simulation manipulating the number of associates to a critical word.

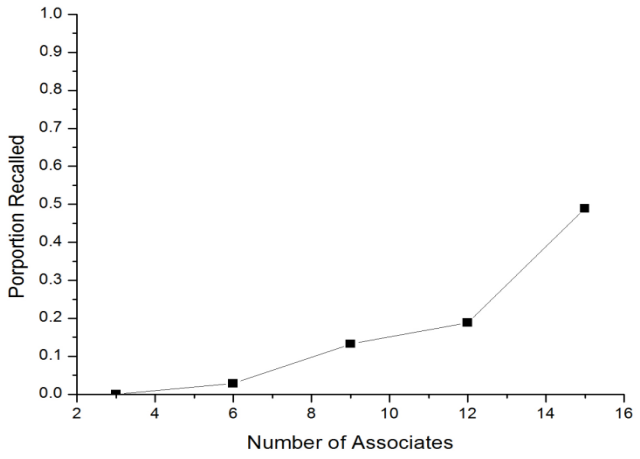

Figure 3. Simulation of Robinson & Roediger (1997).

Simulation #3: Effect of Number of Associates

Robinson and Roediger (1997) have demonstrated that as the number of associates to a critical word contained within a study list is increased, the probability of falsely recalling the critical word increases substantially. Their result strongly suggests that a causal factor underlying false recall is the amount of semantic information about a critical word that is contained in memory. The same pattern should be predicted by our model: as the number of associates to a word is increased, the similarity of the probe to the composite will also increase. This causes the critical lure to have a higher probability of being included in the search set and also to be accepted by the recognition model.

Method We used the same lists as did Robinson and Roediger. (1997). On each repetition, five different DRM lists were selected and 3, 6, 9, 12, or 15 items in the list were randomly selected and added into the study list. Probability of the model recalling a critical word was recorded.

Results The results of this simulation are displayed in Figure 3. This figure shows convincingly that the amount of semantic information contained within the memory vector about a particular word is driving the levels of false recall. As the number of associates to a critical word is increased, the probability of the model falsely recalling the critical word increases as well. In Johns & Jones (2009) we show that the recognition component of our model (RSA) is able to simulate the results of Robinson & Roediger (1997) when this same experimental setup is tested through a recognition experiment, rather than a recall test. Hence, there are two reasons the model is able to account for this result: the search set is more likely to include the critical word as the number of associates to a critical word is increased, and the recognition model is more likely to accept it.

Simulation #4: Categorical vs. Associative Recall

Park, Shobe, & Kihlstrom (2005) examined the levels of false recall that are seen with critical lures for associative lists and categorical lists. Specifically, the categorical study lists were composed of subordinate (vertical) category instances, while the association lists contain horizontal free associates (the typical DRM lists). Park, et al. (2005) found significantly lower levels of false recall for the category labels than for the DRM critical words.

Method The category labels were taken from Park, et al. (2005). The corresponding category lists were attained from the Battigue & Montague (1969) norms and the DRM lists were the six used in the Park, et al. study. Study lists were created by adding four lists from both of these categories. The probability of accepting the critical lure for the DRM lists and the category lists was recorded. In addition, the probability of accepting the studied items was also recorded.

Table 1. Simulation of Park, et al. (2005)

	Category Lists		DRM Lists	
	Old	Critical	Old	Critical
Data	0.79	0.0	0.74	0.33
Model	0.59	0.0	0.675	0.23

Results The results of this simulation are displayed in Table 1. The levels of false recall to DRM lists and category lists for the model are very similar to those found by Park, et al. (2005). That is, the level of false recall to the DRM list is much higher than that for category labels, mirroring what was found in the empirical study using the same materials.

Simulation #5: Relationship between False Recognition and False Recall

Both Stadler, et al. (1999) and Gallo & Roediger (2002) have reported a significant correlation between levels of false recognition and levels of false recall. Stadler, et al. report $r = 0$ .77, $p < 0.001$ across lists, while Gallo & Roediger (2002) report $r = 0.78$, $p < 0.001$ between recall and recognition across their lists. These correlations are likely artificially inflated because in these studies recognition occurred after a recall period, but it does show that there is a relationship between levels of false recall and

recognition. Due to the fact that in our recall model a recognition component is used to make a decision about whether a specific word occurred or not, we expect the model to attain a similar relationship between the predicted levels of false recall and false recognition.

Method Levels of false recall were simulated for the lists contained in Stadler, et al. (1999) and Gallo & Roediger (2002). Levels of false recognition were computed with the RSA model, described in Johns & Jones (2009), for these same lists. 250 trials were done for both models.

Results For the lists from Stadler, et al. (1999) a correlation of $R = 0.578$, $p < 0.001$ was obtained, and for the lists from Gallo & Roediger (2002) a correlation of $r = 0.574$, $p < 0.001$ was found between the levels of false recognition and false recall. These correlations are smaller than those reported in the behavioral results because, as described above, recall preceded recognition in those studies. This relationship is not surprising considering that the RSA recognition model plays an intricate role in our recall model. However, it does demonstrate that the searching mechanism employed by the recall model plays a key role in the levels of false recall. This simulation shows that both the searching and the recognition processes contained in the recall model are creating the level of false recall that the model attains.

Conclusion

This model demonstrates the power of using cognitively plausible representations of words. By incorporating semantic representations based on a co-occurrence learning model, we require only a very simple processing model to explain a wide range of false recall data. In addition, the model is more constrained than models using hand-coded semantic representations because it provides an account of both memory structure and process, and how the two interact to produce false recall.

Further, this approach not only allows us to make quantitative predictions about levels of false recall expected in DRM lists, but it allows for the integration of models that are used to explain different aspects of memory. Integrated together with the SDM model (Johns & Jones, 2008) and the RSA model (Johns & Jones, 2009) these models explain a significant number of effects across many different paradigms. This integration also allows for a greater simplicity across all of the models, as well as the possibility to be combined with other models, such as the TCM (Sederberg, Howard, & Kahana, 2008). Across all three models there are a total of four parameters, which is less than many models designed to explain a single paradigm. There are obviously many results that these models cannot explain, but due to the simplicity of these different models it is an appealing architecture to investigate with other memory phenomena.

References

Battig,W.F., & Montague,W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph, 80*(3, Part 2).

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*, 164-169.

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

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology, 12*, 263-282.

Gallo, D.A.,& Roediger, H.L. (2002). Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring. *Journal of Memory and Language*, *47*, 469-497.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007).Topics in Semantic Representation. *Psychological Review*, *114*, 211-244.

Johns, B. T., & Jones, M. N. (2009). False recognition through semantic amplification. *Proceedings of the 31st Annual Cognitive Science Society.*

Johns, B. T., & Jones, M. N. (2008). Predicting lexical decision and naming times from a semantic space model. *Proceedings of the 30th Annual Cognitive Science Society* (pp. 279-284). Austin, TX: Cognitive Science Society.

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

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review, 114*, 954-993.

Kintsch, W. (1970). Models for free recall and recognition. In D. Norman (Ed.), *Models of Human Memory.* New York: Academic Press, 333-374.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

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

Recchia, G., Johns, B. T., & Jones, M. N.(2008). Context repetition benefits are dependent on context redundancy. *Proceedings of the 30th Cognitive Science Society Meeting*, 267-272.

Robinson, K., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 389-393.

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

Stadler, M.A., Roediger, H.L., & McDermott, K.B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *29*, 424-432.