False Recognition through Semantic Amplification

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

This paper describes a computational model to explain a variety of results in false recognition. The processing mechanism in the model is built around a co-occurrence representation of lexical semantics, affording an account of both structure and process. We show that this model can naturally account for levels of false recognition that are seen in studies using the DRM paradigm, including item-level effects, reaction times, and event-related brain potentials.

Keywords: False recognition; co-occurrence representations; memory models; recognition memory; semantics

Introduction

False recognition is one of the most empirically studied phenomena in recent times, however very little formal modeling of this effect has been conducted. False recognition has been most studied using the Deese/Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). In this task, lists of words that are associated with specific critical words are presented to subjects, and on subsequent memory tests the unpresented critical items are falsely recognized at almost the same level as studied items (Roediger & McDermott, 1995). For example, given *nurse, hospital, sick,* and *cure* to remember, subjects are likely to subsequently produce a false alarm to *doctor.*

Work within the DRM paradigm has provided fundamental evidence about the organization of human memory. The paradigm demonstrates that humans use semantic information to both store and retrieve items, and the use of this information can lead to profound memory errors. However, the exact mechanisms that underlie the false recognition of associates have evaded a formal explanation. Rather, theorists have focused more on general conceptual frameworks of cognition, such as Fuzzy Trace Theory (FTT; Brainerd & Reyna, 2002), the source-monitoring framework (Johnson, Hashtroudi, & Lindsay, 1993), and the discrepancy-attribution hypothesis (Whittlesea, 2002), to explain false memories.

There are now many different computational models of recognition memory. However, a principled problem with these models is that the representations they use do not contain semantic information about specific words, due to the fact that their representations are typically constructed with random number generators. This practice is natural because the models are not typically used to simulate semantic effects in recognition memory. Further, it is still the subject of much debate what are the correct features to represent word meaning. However, we believe that in order to model semantic behaviors, such as is seen with false recognition, one must use a representation of words that contains semantic information. One promising avenue for these semantic representations are those created by co-occurrence learning models.

In a co-occurrence model, a word's semantic representation is constructed by observing statistical regularities in a large corpus of text. These models can account for a variety of different semantic behaviors (for a review see Jones & Mewhort, 2007). Due to the success of co-occurrence models in other domains, it seems natural to assume that they could also be used to account for semantic effects in recognition memory. The models are particularly promising given the observation that associative variables are important predictors of false recognition and false recall, particularly backward association strength (Deese, 1959; Gallo & Roediger, 2002). Because co-occurrence representations correlate with backward association strength (Johns & Jones, 2008), their representations are appealing to be used in a processing model of false recognition. By using a representation of words that is built up through exposure to the environment, we are not simply assuming a particular semantic organization. Instead we are both explaining how a certain memory structure is created, and how this structure interacts with the processing mechanism.

The Recognition through Semantic Amplification (RSA) Model

Our false recognition model is based on the *Iterative Resonance Model* (IRM) of recognition memory (Mewhort & Johns, E., 2005). The motivation for IRM comes from a series of experiments demonstrating that *Old* responses and *New* responses are based upon different types of information (Mewhort & E. Johns, 2000). Note that a subject responds *Old* if the probe item was in the encoded list, and *New* if it was not. In particular, the authors found that the amount of contradictory information contained within a probe predicted *New* responses, whereas *Old* responses were based on the similarity of the probe to the memory items. The original IRM used this dual-criterion decision process. If a decision is not made on a particular information sample, then successive iterations are employed to sharpen the evidence. The number of iterations required for the model to make a decision is taken as a proxy of response latency.

Our *Recognition through Semantic Amplification* (RSA) model is kept within the same formal framework as IRM,

but differs mainly in its representation assumptions. Rather than a set of randomly generated arbitrary features, RSA uses vectors that are built from a co-occurrence learning process. The model assumes that the semantic representations for words presented in a DRM list are retrieved from long-term memory and are stored in a composite memory store (containing a mix of semantic vectors for all the items on the list). When presented with a probe, the model uses a searching (amplification) process to determine whether the probe is similar enough to the composite store to respond *Old*, whilst simultaneously searching for contradictory information between the elementwise comparison of the probe and the composite store to respond *New* (i.e., old and new responses are based on different information, and the response is the winner of a race between the processes). In addition, we demonstrate that the model can be used to simulate both choice probability and response latency results within false recognition by using the number of iterations to make a decision (as in Mewhort & E. Johns, 2005).

The RSA model may be divided into four main components: 1) a co-occurrence representation 2) encoding, 3) amplification, and 4) decision. The processing model works by first encoding all the words that are seen in a specific study list into a single composite vector. This represents the 'gist' of the words that were seen. Then at test, the model attempts to amplify the probe word in this composite vector. If the probe word is in the memory store, its representation should be efficiently amplified. If the probe is not contained in the memory store, it will not be efficiently amplified. Each of these processes will be described in turn. The pseudocode routine for the RSA model is displayed in Figure 1, and the different processes are described formally in this figure.

```
do i=1, number_studied              % encoding process
   memory = memory + (normalize(item(i)) * random)
enddo
probe = normalize(probe)
repeat
 iter = iter + 1
 do i = 1, length_vec               % calculate contradictory
   if (probe(i) > 0)
  cont_info += |norm_to_1(comp(i)) – norm_to_1(probe(i))|
   endif
 enddo
 cont_count += cont_info
 similarity =cosine(probe, memory)/iter    % similarity
 do i = 1, length_vec               %amplification process
    if (probe(i) > 0)
       memory(i) = memory(i) + (probe(i) * (similarity/iter))
   else memory(i) = memory(i) * random
 enddo
 memory = normalize(memory)
until ( (similarity > YES_crit) .or. (cont_count > NO_crit) )
```

Figure 1. Pseudocode listing for the RSA model.

1. Semantic Representations

A word's semantic representation in memory is built using a recent co-occurrence model entitled the Semantic Distinctiveness Memory (SDM) model (Johns & Jones, 2008). The SDM model is a co-occurrence learning model that was created in order to account for the effect of semantic distinctiveness on a word's strength in memory. We have shown, in 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 a better fit to both lexical decision/naming times and semantic organization than classic learning models, and these representations give a good account of semantic isolation effects, semantic similarity ratings, and association norms.

For the purpose of the current paper, the SDM vector representations can be thought of as similar to those created by other co-occurrence learning models, such as LSA (Landauer & Dumais, 1997). A principled difference is that SDM vectors are sparse vectors representing the weighted contexts in which words have co-occurred; this sparsity is optimal for our recognition process borrowed from IRM. The SDM semantic representation for every possible word is stored in long-term memory, and the representations for words presented are retrieved from this store and cast into a short-term store when encoding a DRM list.

2. Encoding Phase

The memory store that the processing model operates on is a single composite vector. Every word presented during the study phase is retrieved from the SDM mental lexicon and is added into this composite vector. Word vectors are first normalized, so that each word adds in approximately the same amount of information. Each vector is multiplied by a uniform random number between 0 and 1 to simulate encoding failure. The composite representation may be thought of as a superposition of items presented in the list.

Other models, such as TODAM (Murdock, 1982), also use a composite vector to create a representation of an event. The different TODAM models use holographic vectors, whereas our vectors are simply summations of semantic traces, but both assume a single vector to create a representation of a study list. The practice is also a similar flavor to the proposal of fuzzy-trace theories, where a 'gist' representation of an episode is created. Even though fuzzy-trace theory seems to entail more sophisticated processes of gist extraction, this encoding phase does correspond with some of the claims of this theory.

3. Amplification Process

The amplification process that the RSA model employs essentially works by attempting to 'turn up the volume' of the probe's representation in memory, while dampening all other items. This causes the probe's representation within

memory to increase across iterations. How strongly the probe is amplified in the composite is determined by a normalized similarity value. This value is determined by taking the cosine between the probe and the memory vector and dividing it by the current iteration. Hence, even though the cosine increases due to the amplification process, the amount of increase is constrained by the current iteration that the model is in. If a decision is not made on the current iteration, the amplification process is repeated.

The second mechanism works by iteratively dampening the non-defining elements of the probe in the composite. This is accomplished by simply multiplying the memory vector by a random number between 0 and 1 at each location where the probe word contains no information (i.e. where the probe vector is 0). With this process, a word that occurred in the study list is amplified more efficiently within the composite because it contains more semantic information and also contains less contradictory information that needs to be filtered out of the composite.

4. Decision Process

The model uses two different sources of information to make a decision about whether to accept or reject a probe. As in the IRM, *Old* responses will be based on the similarity of the probe to the memory vector, while *New* responses will be based on the amount of contradictory information that the probe did not occur in the study list. The similarity value will be assessed with a cosine between the probe and the memory vector. If this similarity value exceeds a certain criterion then the probe is accepted. In the following simulation, this yes criterion is set at 0.99.

The amount of contradictory information is assessed by measuring the difference in the pattern of the probe and the memory vector. This is computed as the absolute difference between the defining portions of the probe and the corresponding locations within the memory vector, when both of the vectors are normalized to have a magnitude of 1. This returns a value between 0 and 1, which 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. Since the amount of contradictory information will decrease across iterations (due to the amplification process), the amount of contradictory information is a running count. If this count exceeds a certain criterion, then the probe is rejected. In the following simulations this *No* criterion is set to 4.5. Thus, contradictory evidence is the difference in the values that define the probe word's semantic representation (the non-zero entries in the word's episodic trace) and the corresponding values in the composite vector.

Discussion of RSA

The model that we propose here is based on a simple representation assumption – that all the words seen in a study list are added into a single composite vector, or in other words, the composite contains the 'gist' of the study list. To determine whether a word occurred in a specific list, a simple mechanism is employed where a word's representation is amplified within the composite. *Old* decisions are based on a similarity value between the probe word and the memory vector, whilst *New* decision are based on the amount of evidence that the word did not occur. The amplification process essentially works by attempting to filter the probe out of the composite representation, emphasizing signal (if present) while simultaneously dampening noise. If either signal or noise is strong enough, a confirmation/contradiction decision can be made.

This model should be particularly effective at explaining false memory paradigms because the searching mechanism is dependent on the amount of semantic information contained within the composite vector shared with the probe. When the composite contains a large amount of information about a word, then it is amplified efficiently increasing the likelihood of reaching the *Old* criterion than the *New* criterion. Hence, even though a word was not contained in the study list, it could be accepted with a high probability if it shares semantic information with study words. There is very little complexity built into processing model; instead the main locus is put on the contents of memory. To this point, there are only two fixed parameters (both decision parameters) that drive this model.

Simulations

The methodology that we use in simulating false recognition results is very simple: for the exact words used in an experiment we retrieve the semantic representations learned by the SDM model and encode these words into a composite vector representing the list. The two parameters of the RSA model are fixed across simulations. Instead, the locus of memory effects are dependent on the different words in memory, not different processing for different experiments.

Simulation #1: Levels of False Recognition

We first test the RSA model on is whether it attains similar levels of false recognition to those seen in behavioral data. We will simulate three different sets of DRM lists: 1) the lists from Roediger and 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).

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. The model was then tested with studied items and critical lures. Average hit and false recognition rates were recorded across 1000 trials. To test levels of false recognition to unrelated items, five words were randomly selected from the Toronto Word Pool (Friendly, et al., 1982), for each trial.

Results The levels of recognition for studied, critical, and unrelated words for the model and each of the studies are

displayed in Figure 2. This figure shows that the model gives a very good approximation to the qualitative trends seen in the behavioral data across the different word types. Hence, the RSA model seems to be susceptible to the same type of memory illusions that humans are. This is because the model amplifies the critical word trace efficiently due to the memory vector containing a large amount of semantic information about the meaning of that word. This figure also shows that the model has close approximation to the level of recognition for unrelated lures.

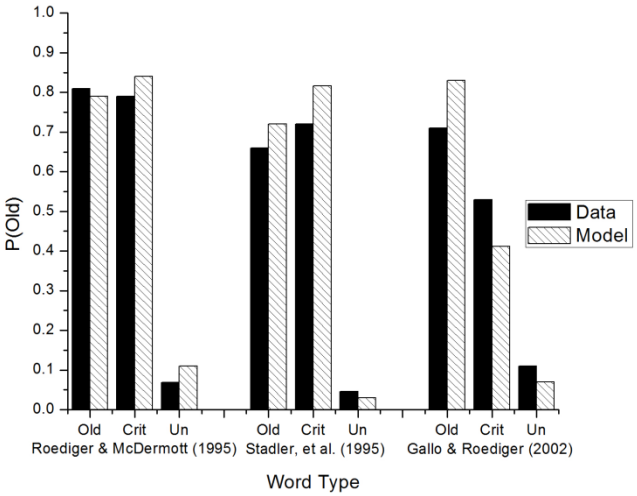

Figure 2. RSA levels of true and false recognition.

Simulation #2: Item-Level Analyses

Stadler, Roediger, & McDermott (1999) and Gallo & Roediger (2002) both published the levels of false recognition that are seen with different DRM lists. As both of these studies show, there is considerable variability in the levels of false recognition elicited by different DRM lists. To test the model's quantitative predictions of false recognition, we correlated the levels of false recognition for both the model and the behavioral data using the same words reported in behavioral data. This allows us to be confident that both the memory structure that the model is utilizing, and the processing mechanism that is operating on these representations, are working together to produce false recognition in a manner coherent with experimental data.

Method Levels of false recognition were attained from Stadler, et al., (1999) and Gallo & Roediger (2002). Again, four DRM lists from the different studies were added to form a single study list. The level of false recognition for each critical word was recorded across 1000 simulated trials. In addition, we computed the raw cosine between the critical word vector and the composite vector for each critical word (i.e., how much variance is predicted by solely by the structural representations without the process mechanism). This allows us to test the respective contribution of the memory structure and the process in creating false recognition.

Results Across the 55 critical words (with repeats removed), a significant correlation of $r = 0.476$, $p < 0.001$ was obtained between the data and predictions of the model. If the five lists that the model does worst on are removed (*thief, needle, king, trash*, and *car*), the correlation increases to $r = 0.649$, $p < 0.001$. To assess what impact the memory structure is having on false recognition, the cosine was computed between the composite list vector and the individual word vector for each list. A correlation of $r = 0.333$, $p < 0.05$ was obtained between the level of false recognition and cosine across the 54 lists. This demonstrates the semantic representation of words has sufficient power within it to predict item-level amounts of false recognition. When combined with a simple process mechanism that is designed to exploit word structure, we see a better fit to the data than either structure or process alone can accomplish. Hence, it is the interaction between the structure of memory and the process mechanism that produces the superior fit to the data, not simply the structure or process alone.

Simulation #3: Effect of the 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, a systematic increase in false recognition rates to that critical item is also observed. This is an interesting study for the RSA model to simulate because it suggests that the number of semantic associates is the causal factor in determining the false recognition rates of a critical item. Hence, we expect a similar pattern of results to Robinson and Roediger because the more semantic associates that are studied, the more efficiently a critical lure should be processed, which in turn should lead to an increased hit rate for these items.

Method We used the same lists as Robinson, et al. (1999). On each repetition, five DRM lists were selected and 3, 6, 9, 12, or 15 items from the list were randomly selected and added into the current study list. The levels of false recognition for the critical lures at the different level of associates, as well as the hit rates for the studied items, were recorded across 1000 replications.

Results Figure 3 shows the human data and simulated results for both the hit rate and the false alarm rate for critical words from the Robinson, et al. (1999) study. This figure shows that as the number of associates to a critical word in a study list is increased, the false alarm rate to the critical lure is also increased. The model produces slightly more false alarms (especially with 15 associates), but the same general trend is observed. This simulation demonstrates that high levels of false recognition are seen with the RSA model due to the increased amount of semantic information contained within a study list's episodic representation.

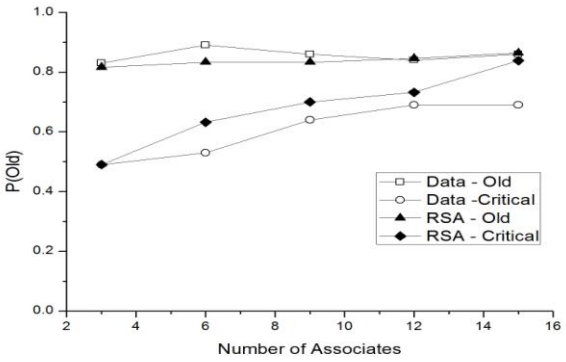

Figure 3. RSA simulation of Robinson & Roediger (1997).

Simulation #4: Effect of Number of Associates on Short-Term Recognition RT and Accuracy

Coane, McBride, Raulerson, & Jordan (2007) conducted a short-term recognition experiment in which they created set sizes of 3, 5, and 7 by sampling from a single DRM list. Reaction time and accuracy were recorded as a function of the number of associates within a list. They found a greater RT for critical lures than studied items, and also an increase in RT as a function of set size for both word types. In addition, they found that the proportion of false alarms to the critical lure increases as a function of set size, showing that there is a false recognition effect even at small list sizes. A compelling feature of the RSA model is that it provides a framework to account for both choice probabilities and reaction time, making this an attractive study to simulate.

Method As in Coane, et al. (2007), set sizes of 3, 5, and 7 were created by sampling from DRM lists from the Sadler, et al. (1999) set of DRM lists. Reaction time was assessed by taking the number of iterations to accept or reject a probe as a proxy of RT. Accuracy for both studied items and critical lures was assessed for the three set sizes.

Results Figure 4 displays the results for both simulated reaction time (top panels) and choice probability (bottom panels) as a function of set size, for both the RSA model (left panels) and the data from Coane, et al. (2007; right panels). As can be seen from this figure the model gives an excellent approximation for both reaction time and choice probability data as a function of set size. The reason the model can capture the reaction time data is that even though the composite contains a considerable amount of semantic information about the critical word, this word is still not as similar as a studied word to the memory store. This means that the word is not amplified as efficiently, so it takes longer for the composite to become similar enough to be accepted or to accumulate enough contradictory information to reject the probe. An increase in set size slows responses to studied items more than the critical word because the unique information for a presented item becomes mixed in with a greater amount of information from other items.

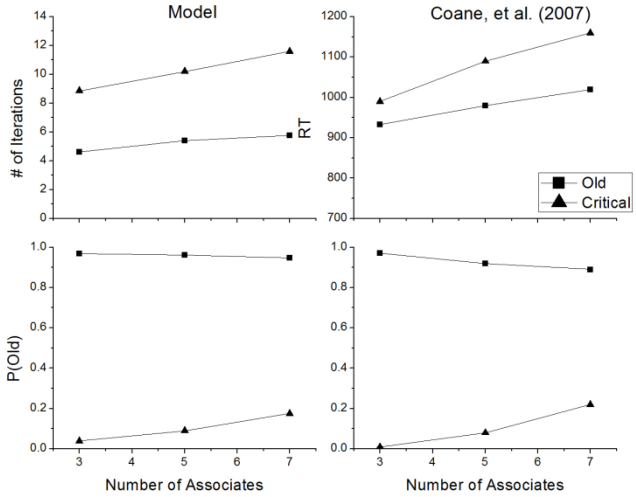

Figure 4. RSA simulation of Coane, et al. (2007).

Simulation #5: ERP Patterns

In order to test whether the process that the RSA employs is cognitively plausible, we compared the change in activation across iterations in RSA with ERP studies examining recognition memory. We do not wish to attempt to localize this process or propose that there is a specific neurological mechanism of this model; instead we simply test whether the model's temporal dynamics change in a similar manner as neurological processes seem to. There are two main results that we would like to focus on.

The first ERP result that we wish to test is the 'N400 strength effect' described in Finnegan, Humphreys, Dennis, & Geffen (2002). In this study, subjects made old/new recognition judgments to strong words (presented three times), weak words (presented one time), and new words. The authors found a greater N400 wave for strong items vs. weak items, and for weak items vs. new items. The second ERP result that we would like to simulate is a study by Johnson, Nolde, Mather, Kounios, Schacter, & Curran (1997). In this study, the experimenters monitored subject's ERP response during old/new recognition decisions when tested in the DRM paradigm. They found a greater waveform for studied words vs. critical words, and for critical words vs. unrelated new words. The value we use to predict activation in the RSA model will be the normalized activation value used in the amplification process.

Method In order to simulate the results of Finnegan, et al. (2002), list sizes of 50 were created by sampling randomly from words within the Toronto Word Pool (Friendly, et al., 1982). In this list, twenty five words were encoded three times (the 'strong' words), while 25 words were only encoded once (the 'weak' words). Then the activation levels for both the strong, weak, and new words (attained by randomly sampling from the word pool) were recorded across iterations.

To simulate the results of Johnson, et al. (1997), DRM lists of size four were created by randomly sampling lists from the Stadler, et al. (1999) study. The activations for studied items, critical words, and unrelated words (obtained by randomly sampling 4 critical words whose list did not occur in the study list) were computed across iterations.

Results Figure 5 displays the simulation results for both Finnegan, et al. (2002) (top panel), and Johnson, et al. (1997) (bottom panel). In the simulation of Finnegan, et al. the model produces a higher activation across iterations for strong vs. weak words, and weak vs. new words, similar to the pattern seen in their study. Strong words are more easily amplified within the composite vector compared with weak words, and weak presented words are more easily amplified compared with unpresented words. In addition, the time course of activation change in the model qualitatively mimics the N400 pattern found by Finnegan et al.

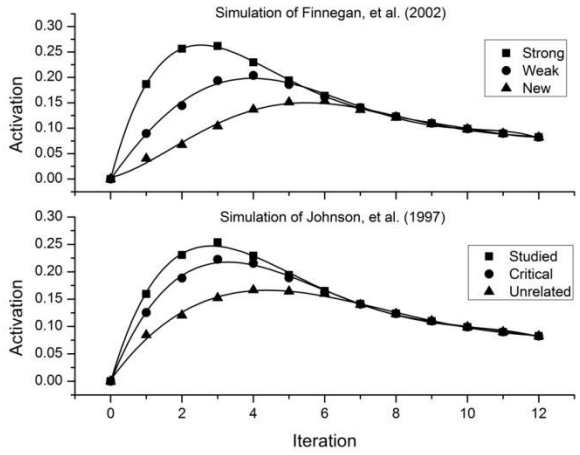

Figure 5. ERP simulations with the RSA model.

We see a similar pattern of results in the simulation of Johnson, et al. (1997), where studied words have the highest level of activation, but the activation levels for the critical lures are also quite high compared with unrelated lures. This demonstrates why the model is able to attain false recognition: a greater amount of semantic information about the critical lure is contained within memory, so it is amplified more efficiently, which in turn makes it easier to accept. As with the previous simulation, the dynamics of activation change over time nicely match the qualitative waveforms presented in Johnson et al. (1997).

Conclusion

We have described a new model of false recognition built on a representation that has proven useful at accounting for other elusive effects in memory. Our RSA model is successful at simulating a range of false recognition data because it fuses a simple process account with a structural representation for words generated from experience with environmental regularities. Critical lures are more likely to share semantic information with studied words and, hence, they are more efficiently amplified in the composite memory store. Our model leaves much of the complexity required to produce false recognition behavior in the semantic representations (learned from language), allowing it to use a much simpler processing mechanism, and without reliance on hand-coded word representations.

References

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

Coane, J. H., McBride, D. M., Raulerson, B. A., & Jordan, J. S. (2007). False Memory in a Short-Term Memory Task. *Experimental Psychology*, 54, 62-70.

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

Finnegan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP 'old/new' effects: memory strength and decisional factor(s). *Neuropsychologia*, 40, 2288-2304.

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.

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

Johnson, M. K., Nolde, S. F., Mather, M., Kounios, J., Schacter, D. L., & Curran, T. (1997). The similarity of brain activity associated with true and false recognition memory depends on test format. *Psychological Science*, 8, 260-257.

Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory*, 13, 300-307.

Mewhort, D. J. K., & Johns, E. E. (2000). The extra-list feature effect: A test of item matching in short-term recognition memory. *JEP: General*, 129, 262-284.

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

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. *JEP:LMC*, 21, 803-814.

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

Whittlesea, B. W. A. (2002). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *JEP: General*, 131, 96-115.