The Inverse List Length Effect: Implications for Exemplar Models of Recognition Memory

Simon Dennis (simon.dennis@gmail.com)
225 Psychology Building, 1835 Neil Avenue
Columbus, OH 43210 USA

Allison Chapman (achapman.274@gmail.com)
225 Psychology Building, 1835 Neil Avenue
Columbus, OH 43210 USA

Abstract

A. H. Criss and R. M. Shiffrin (2004) argued against the composite context noise explanation of recognition memory introduced by Dennis and Humphreys (2001) by showing that with novel face stimuli, related distractors that are drawn from study categories show elevated false alarm rates. Dennis and Humphreys (2001) proposed two mechanisms by which related false alarm rates might arise. The first mechanism posited that such items might be produced as implicit associative responses. Such a mechanism is unlikely to apply to face stimuli. The second mechanism posited a category wide criterion shift, which Criss and Shiffrin (2004) argued was implausible because it requires criterion to be adjusted on an item by item basis during test. Rather they contended that direct interference between study items was more likely to account for the data and fit the Retrieving Effectively from Memory model (REM, Shiffrin & Steyvers, 1997) to their data. We suggest that the mechanism by which false alarms are generated for word and novel face stimuli may differ. Instead of focusing on related distractors, we focus on unrelated distractors. Using words, we show that if list length is manipulated by keeping the number of categories constant but increasing the number of exemplars in each category, then the unrelated false alarm rate decreases - thus inducing an inverse list length effect in which performance on short lists is worse than on long lists. Simulations demonstrate that exemplar models such as REM are not capable of accounting for the results without modification. Rather a composite representation of the studied categories that subjects can use to reject unrelated lures must be formed.

Keywords: recognition memory, category effects, list length effect

Introduction

Criss and Shiffrin (2004) in a comment on Dennis and Humphreys (2001) argue that episodic recognition memory involves both item noise generated by the other items that appeared during the study episode and context noise generated by other contexts in which a test item has been seen. They present the results of two experiments. The second is a multilist design which demonstrates that at least under some conditions context noise seems to play a substantial role. This result supports the main point of Dennis and Humphreys (2001) and we will not discuss it further.

The first experiment involved the presentation of word/face pairs at study. The words were drawn from both semantic and orthographic/phonological categories and face stimuli from face categories defined by gender, race and age. The length of these categories was manipulated independently at study (2, 6 or 9 items). At test, words and faces were presented in separate blocks and participants were required to make two judgements. Firstly, they were required to rate how confident they were that the item appeared on the study list. Secondly, they had to indicate how many similar items appeared at study. The probability of correctly identifying studied items as studied increased with category length as did the probability of incorrectly identifying related distractors. Furthermore, participants' estimates of the number of related items at test increased with the actual category length.

Criss and Shiffrin (2004) proposed a version of the Retrieving Effectively from Memory model (REM, Shiffrin & Steyvers, 1997) that incorporated both item and context noise. Their account of the category length effects drew on previous work on the global matching models (Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989) which attributes category effects to the overlap in item representations. Related lures are assumed to overlap with many studied items and as a consequence have higher match values. The more category exemplars on the list the higher the match value is likely to be and the more likely it is that a false alarm will be generated.

By contrast, Dennis and Humphreys (2001) argued that similarity effects like those found by (Criss & Shiffrin, 2004) could be captured in a context noise model in at least two ways. The first proposed that when exposed to categorically organized lists of this kind, subjects might generate implicit associative responses (IARs, c.f. Underwood, 1978). IARs are self generated items that are bound to the study context not because they actually appeared, but because they were internally generated by the subject. For instance, if the subject received a list containing the words, "bed", "night", "dark", ..., then they might think of the word "sleep" and in so doing establish an association to the list context. Subsequently, the existence of this association might increase the likelihood of incorrectly recognizing "sleep". Criss and Shiffrin (2004) argued that such a mechanism was implausible with face stimuli because it would not be clear what could be generated as the IAR.

The second mechanism proposed by Dennis and Humphreys (2001) was a category wide criterion shift. If participants were conscious of the categories in the list they might be inclined to use a low criterion for studied categories because they knew the category had been studied and hence might use this information to inform their decision even in the absence of item information. Criss and Shiffrin (2004) argued that this was implausible, citing work by Wixted and Stretch (2000) that has shown that participants are sometimes reluctant to adjust criterion on an item by item basis during test.

We agree that IARs are an unlikely explanation for similarity effects with face stimuli, however, it still seems plausible that subjects were adjusting criterion on an item by item basis during test. More recently it has been demonstrated that item by item criterion shifts occur (Singer & Wixted, 2006) and we do not believe it is possible to rule out the possibility *a priori*. Furthermore, the fact that subjects were able to give somewhat accurate estimates of category length on an item by item basis at test suggests that they had exactly the kind of information necessary to manipulate criterion. The instruction to think about the category that the item belonged to in order to complete the second component of each question would seem to make that information even more salient than it might otherwise have been. So, we would not concur that this experiment demonstrates the necessity of item noise as the explanatory mechanism of similarity effects.

Furthermore, it is not necessarily the case that the substantive mechanisms responsible for false alarms to related lures with word stimuli are the same as those for novel face stimuli. The words typically used in recognition memory experiments are well learned and as a consequence the corresponding representations may overlap little. By contrast the novel face stimuli may be similar to each other on a number of dimensions. There are two main sets of data that suggest that this may be the case. Firstly, list length effects seem to be present for novel face stimuli, but when appropriate controls are employed are not in evidence with words (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, submitted). Secondly, in list strength paradigms in which some stimuli are strengthened through additional presentation time, or more presentations, the presence of strong words does not affect performance on the weak words (Ratcliff, Clark, & Shiffrin, 1990). However, with novel face stimuli this is not the case (Norman, Tepe, Nyhus, & Curran, 2008). These results as a set might indicate that word stimuli overlap little, while novel face stimuli overlap to a greater degree.

It does not seem possible to resolve the issue by focusing exclusively on false alarms to related lures. Instead, we intend to focus on false alarm rates to unrelated lures when list length is manipulated by drawing different numbers of exemplars from a fixed number of taxonomic categories. Exemplar models like REM predict that as list length increases it should become increasingly difficult to reject unrelated items as the amount of item noise increases. This result occurs because every additional item in the study list, even if it is from a different category, is assumed to add variance to the familiarity calculation and consequently to decrease discriminability.

However, if subjects are forming a representation of the studied categories that is separate from their representations of the list exemplars then they could compare a test item to this composite representation to determine if the test item appeared in a studied category. Under these conditions, one might expect that an item that did not appear in any of the categories - a unrelated lure - would induce a higher criterion and tend to be rejected. As the number of exemplars from the category increased and the subject becomes increasingly confident about the exact nature of the categories that appeared at study, one would expect the false alarm rate to unrelated lures to fall.

We start by presenting two experiments in which list length was manipulated as indicated above. We then present simulations showing that the REM model is not capable of accounting for the data without modification.

Experiment One

In the first experiment, we drew stimuli from eight taxonomic categories and blocked the presentation of exemplars from a given category to encourage categorical representation during study.

Method

Participants Participants were thirty, undergraduate students enrolled in an introductory psychology course at Ohio State University. Course credit served as incentive for participation.

Materials Stimuli included words selected from the Overschelde, Rawson, and Dunlosky (2004) category norms. Data concerning response proportions and response latencies were examined to select eleven exemplar words commonly associated with each of forty-eight categories. The 528 total words were reported 30% of the time, on average. Mindful of semantic ambiguity, we strove to exclude words with multiple meanings and also words nearly identical in connotation.

Procedure Three list lengths – short, medium, and long – were presented within-subjects across three study-test cycles. The short, medium, and long study lists contained eight, twenty-four, and eighty words, respectively. Each study item was presented for three seconds. Test lists were twenty-four words in length, including: eight target words (randomly selected from each category in the study list); eight related distracters (one unstudied exemplar from each study category); and eight unrelated distracters (one from each of eight categories not otherwise present on any study or test list). Each probe word remained on the screen until participants registered a yes/no recognition judgment.

The experiment was counterbalanced for order such that there were six total conditions, containing five participants each. Retention interval was equated retroactively wherein a puzzle appeared 696 s following short-, 648 s following medium-, and 480 s following long study lists. The effects of retention interval were thereby controlled. Departing from standard retroactive design, target words were drawn from both the start and end of study lists. Exemplar words were shuffled within but not between categories prior to study list presentation; contrastingly, test list target exemplars and lures were presented randomly across categories.

Subjects were asked to attend to items in each study list, attempt to solve the subsequent puzzle test, and then complete a word recognition test by pressing P (Old) or Q (New). Instruction did not divulge information regarding the disparate list lengths nor the categorical nature of the word stimuli. Furthermore, no indication was made to assert that the recognition test was more important than the puzzle test. Before the test list began, reminder instructions were given.

Results

The results are summarized in Figure 1. An ANOVA of hit rate (HR) revealed no statistically significant difference between list lengths, F(2, 28) = 0.948, p>.05. Interestingly, tests of both related lures and unrelated lures yielded false alarm rates (FARs) that were significantly different in short- versus medium- versus long-lists. As expected, the related FARs differ as a consequence of list length, F(2, 28) = 6.49, p<.01. The means in this case follow from the assumption that related false alarms increase as study item presentation increases: 0.256 for short, 0.246 for medium, and 0.389 for long. Critically, the ANOVA of unrelated FARs indicates a statistically significant result: F(2, 28) = 4.716, p<.05. As is evident in Figure 1, the mean FARs for the unrelated lures tend to decrease as the study list lengthens, with mean FARs of 0.224 (short), 0.199 (medium), and 0.145 (long).

Experiment Two

In the second experiment, we investigated the extent to which the pattern of results obtained in the previous experiment relied on blocking items from the same category at study by distributing the items from a given category through the list. It may be that when the category structure is made less obvious as is the case in many experiments on similarity effects, that subjects will be unable to use category information to exclude unrelated distractors. The procedure was identical to that outlined above in all other respects.

Results

The results of Experiment 2 parallel the results of the first experiment, although the distributed nature of categorical presentation appears to reduce the effects observed in Experiment 1. Furthermore, there seems to be a criterion shift in the short condition (see Figure 2). The analysis of hit rate variance suggests no difference between the short, medium, and long lists where F(2, 28) = 1.271, p>.05. When examining the FARs for related distracters, we find a significant difference between conditions with F(2, 28) = 5.188, p<.01. The mean false alarm rates in this instance appear to increase with length: 0.152 for short, 0.222 for medium, and 0.259 for long. The means for unrelated FARs were not significantly different, F(2, 28) = 1.557, p>.05. These are the means for the unrelated FARs in distributed: 0.141 for short, 0.182 for medium, and 0.137 for long. Note that the false alarm rate decreases from the medium to the long list.

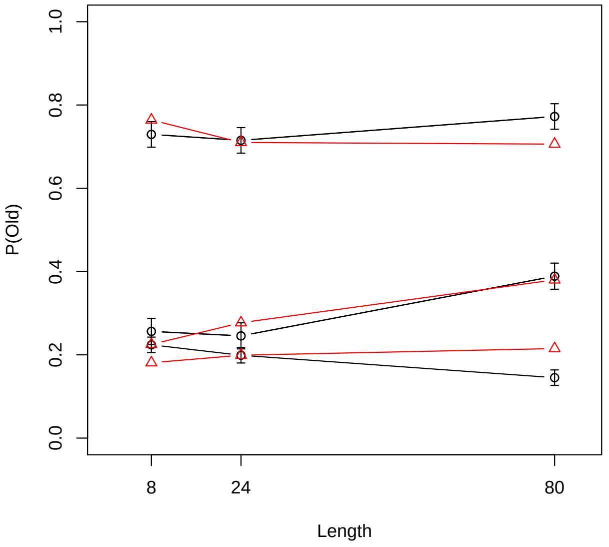

Figure 1: The probability of saying yes in experiment one as a function of list length (8, 24 or 80). From top to bottom the lines represent the hit rates, false alarms to related lures and false alarms to unrelated lures. Error bars indicate the standard errors. Triangles indicate the fit of the REM model to the data.

Discussion

The Global Matching Models of recognition memory (Gillund & Shiffrin, 1984; Murdock, 1982; Hintzman, 1984; Humphreys, Bain, & Pike, 1989; Clark & Gronlund, 1996) propose that when items from a given category are present on a list, related lures will have higher false alarm rates because the representation of the lure overlaps with the representations of each of the category exemplars from the list and therefore will tend to have a higher match value which will be more likely to exceed criterion.

The representational overlap explanation of category effects is elegant and may well play an important role in category effects for stimuli such as novel faces. As suggested above, however, the words that are typically used in recognition memory experiments are well learned. In order for verbal retrieval to be generally effective, words must be well discriminated and list length results suggest that overlap is negligible (Dennis & Humphreys, 2001; Dennis et al., 2008).

By manipulating list length as we have and focusing on unrelated lures we pit the noise properties of items against the category effects. Increased list length tends to increase item noise and compromise performance. Additional exemplars should improve the category representation of the list that the

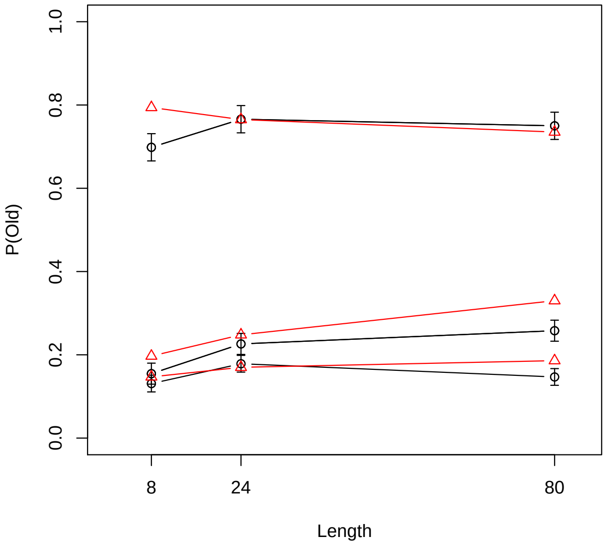

Figure 2: The probability of saying yes in experiment two as a function of list length (8, 24 or 80). From top to bottom the lines represent the hit rates, false alarms to related lures and false alarms to unrelated lures. Error bars indicate the standard errors. Triangles indicate the fit of the REM model to the data.

subject is forming and make it easier for them to exclude the unrelated lures. As we saw in the results section, performance improves with list length and in particular the false alarm rate to unrelated lures decreases.

To check our intuitions about the item noise models, we chose to fit the REM model to the results of experiment one. REM incorporates the overlap explanation of category effects common to the global matching models. Furthermore, it does so by proposing separate storage of item exemplars. False alarms to related lures occur as a consequence of summing matching strength from separate comparisons of each of the exemplars to the lure item rather than positing that a composite representation of the studied categories is accumulated as the items are studied. We will start by giving a brief description of the model and then outline our simulations. Readers should consult the original paper (Shiffrin & Steyvers, 1997) for a more thorough treatment of the model.

Retrieving Effectively from Memory (REM)

According to the REM model, encoding involves copying features from a lexical/semantic representation of an item into a corresponding episodic vector. The features in the traces are integers drawn from a geometric distribution. The probability of drawing feature ν is:

$$P(\nu) = g(1-g)^{\nu-1}, \nu > 0 \tag{1}$$

The g parameter of the distribution determines the likelihood of generating larger features and is assumed to change with characteristics such as the frequency of the word. The copying process is noisy and so features are copied with a probability c and on some occasions are incorrectly copied with probability u. If a feature is incorrectly copied its value is determined by an independent draw from the geometric distribution. Once a feature is stored it does not change during an experiment.

At retrieval, the representation of the test item is compared to each of the items that appeared at study (in this version of the model it is assumed that context has already isolated these items) and a likelihood is calculated according to the following equation:

$$\lambda_{(i,j,k)} = (1-c)^{nq_{(i,j,k)}} \prod_{\nu=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{\nu-1}}{g(1-g)^{\nu-1}} \right]^{nm_{(\nu,i,j,k)}} \tag{2}$$

where $nq_{(i,j,k)}$ is the number of nonzero features that mismatch for a study item i, test item j and subject k and $nm_{(\nu,i,j,k)}$ is the number of matches of the feature ν. The overall odds that a given test item is a studied word is determined by averaging these likelihoods:

$$\Phi_{jk} = \frac{P(old|features)}{P(new|features)} = \frac{1}{N}\sum_{i=1}^{N} \lambda_{(i,j,k)} \tag{3}$$

The natural criterion for the system is at 1.0. Likelihoods above 1.0 imply that the probability that the word is old is greater than the probability that the word is new.

To capture the notion of overlap, a representation of the category was first drawn from the geometric distribution. Then each of the items belonging to that category were created by independently changing the features with probability o. If a feature was changed it was redrawn from the geometric distribution.

Simulations

A simplex method was used to fit the model (Nelder & Mead, 1965). At each point 1000 subjects were simulated to ensure accurate estimates of the hits and false alarm rates. The number of features was set at 20. Figure 1 shows the fit of the model to the data. Parameters were c = 0.83, u = 0.17, g = 0.30, o = 0.41. Sum of squared error = 0.014. As anticipated the hit rates that the model predicts decrease and most critically the false alarm rates to the unrelated lures increase instead of decreasing as occurs in the data.

Allowing Criterion to Vary

In likelihood models such as REM, it is common to assume that criterion is set at 1.0 on the odds ratio scale (0.0 on the log odds scale). Since many experiments have roughly equal numbers of targets and distractors, it is appropriate to assume that the prior probabilities are approximately equal. This assumption was made in the simulations reported above. It may be, however, that subjects are inclined to a stricter criterion in the case of the long list, where they maybe more concerned about the possibility of a false alarm.

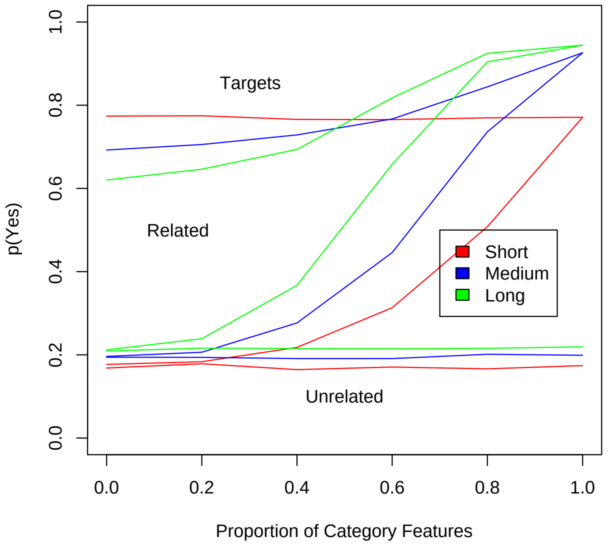

Figure 3: REM simulations showing performance as the proportion of category features increases.

Figure 3 shows performance as a function of the proportion of category features. The first observation is that the proportion has no impact on the false alarm rates to unrelated lures. The false alarms in the long case remain higher. However, when the proportion is above about 0.5 then long hit rates and false alarm rates are higher than medium and short rates and so a global criterion shift could plausibly capture the results.

Figure 4 shows the simulations of the REM model when criterion is allowed to vary freely. Parameters were c = 0.71, u = 0.27, g = 0.37, o = 0.405, short criterion = 1.11, medium criterion = 1.49, long criterion = 1.06. Summed squared error = 0.015. Allowing criteria to vary makes little difference to the performance. In fact, the long criterion is actually set lower than the medium and short, rather than higher as it would need to be to account for the unrelated false alarm rates. The problem is the false alarm rates to related lures. If the proportion of category features is high enough that the long hit rate is greater than the short hit rate then the false alarms to related lures are far too high.

Because of the criterion shift apparent for the short lists in experiment two, it is necessary to posit that the criterion are free in order to capture these results well. Figure 2 shows the fit of the REM model. The parameters were c = 0.69, u = 0.31, g = 0.33, o = 0.40, short criterion = 2.01, medium criterion = 1.25, long criterion = 1.20. Although the short criterion is now higher leading to a reduced false alarm rate in the short condition as seen in the data, the model fails to capture the downward trend between the medium and long conditions.

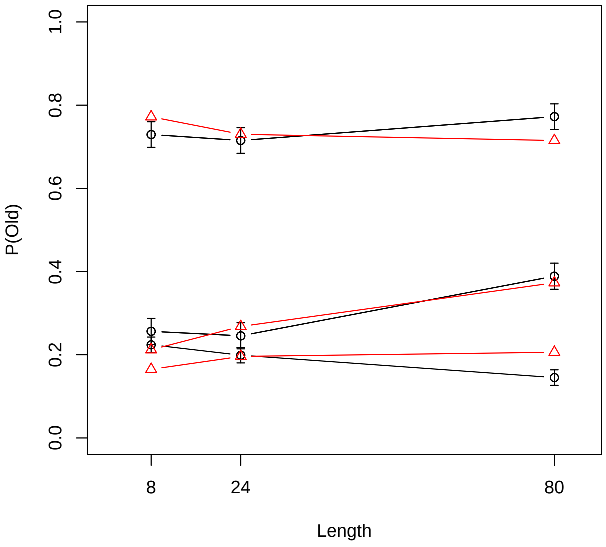

Figure 4: Fit of the REM model to the data from experiment one with criteria free to vary.

Conclusions

When the global matching models were first created, the notion that test items were matched against all of the items from the study list was proposed for two main reasons (Clark & Gronlund, 1996). At the time it was widely accepted that as list length increases performance decreases. There is now significant evidence that the decrements that are found for word stimuli are not a consequence of direct interference between items, but rather occur as a consequence of a number of confounds such as retention interval, attention, rehearsal and contextual reinstatement (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, submitted).

The second main reason was the elegant explanation of category effects. If strength is accumulated over item matches then an item that partially matches many items in a list might induce a false positive response. For stimuli other than words that are not as well learned, like novel faces, this may still be an accurate description. The list length and list strength effects that have been found with these stimuli suggest that the amount of overlap is greater (Kinnell & Dennis, in prep; Norman et al., 2008). However, for word stimuli the current results suggest that something else is going on. An explanation that relies solely on the accumulation of item matches cannot explain why people are better able to exclude items from unstudied categories as list length increases.

Rather we contend that one must form a representation of the categories that appeared on the list - perhaps incorporating this into the contextual representation of the list. As the subject sees more items from the category this representation must become more specific and more effective at excluding the items from unstudied categories. Such an assumption is inconsistent with purely exemplar models such as the existing versions of the REM model. However, one could modify the model in a straightforward way to incorporate an exclusion mechanism.

Acknowledgments

We would like to thank Amy Criss and Richard Shiffrin for helpful discussions about these experiments and the simulations of the REM model.

References

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., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on dennis and humphreys (2001). *Psychological Review*, *111*, 800-807.

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.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1-67.

Hintzman, D. L. (1984). Minerva-2 - a simulation-model of human-memory. *Behavior Research Methods Instruments & Computers*, *16*(2), 96-101.

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*(1), 36-67.

Kinnell, A., & Dennis, S. (in prep). The role of stimuli type in list length effects in recognition memory.

Kinnell, A., & Dennis, S. (submitted). The list length effect: An analysis of potential confounds.

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

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308-313.

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.

Overschelde, J. P. V., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the battig and montague (1969) norms. *Journal of Memory and Language*, *50*, 289-335.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect .1. data and discussion. *Journal of Experimental Psychology-Learning Memory and Cognition*, *16*(2), 163-178.

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

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

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

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 368-376.