What are the boundary conditions of differentiation in episodic memory?

Adam F. Osth (adamosth@gmail.com)
The Ohio State University
1835 Lazenby Hall
Columbus, OH 43210, USA

Simon Dennis (simon.dennis@gmail.com)
The Ohio State University
1835 Lazenby Hall
Columbus, OH 43210, USA

Abstract

One of the critical findings in recognition memory is the null list-strength effect (LSE), which states that repeating study items does not hurt the performance of other studied items. Episodic memory models were able to predict the null LSE by using the principle of differentiation, in which repetitions of an item accumulate into a single strong memory trace. A hypothesized boundary of differentiation is that repetitions of an item in different contexts will create new traces. Two experiments tested this hypothesis by repeating words across different study-test cycles rather than within a single list followed by a test on all of the studied lists. Results indicated that as the proportion of strong items increased, there was both a null LSE and a non-significant decrease in the FAR, which is contrary to the predicted strength-based mirror effect. These two results in tandem provide a challenge for differentiation models.

Keywords: Recognition memory; Episodic memory; Memory models; List-strength effect

Introduction

Episodic recognition memory is typically tested using the yes/no recognition task, in which participants are presented with a list of words to study and during a subsequent test phase are instructed to say "yes" to the previously studied words and "no" otherwise. In the 1980's, a class of simulation models referred to as global memory models were developed to account for data from this paradigm, as well as other paradigms such as free recall, cued recall, and associative recognition. While these models posited different storage operations, such as composite storage, in which all memory traces compile into a single memory representation (TODAM: Murdock, 1982; the Matrix model: Pike, 1984), and multiple trace storage, in which individual memory traces are separable and accessible (SAM: Gillund & Shiffrin, 1984; Minerva 2: Hintzman, 1988), all of these models posit a retrieval process in which the probe item is matched against all of the memory traces from the study list in parallel, producing a "familiarity" value for the test item that is compared to a criterion to make a decision.

A success of the global memory models was the ability to account for the *list length effect*, in which increases in the size of the study list produce decreases in discriminability, an effect which was believed to be a very robust finding in recognition memory at the time (although recent examinations of the list length paradigm have uncovered several confounds, see Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008). These models were able to make this prediction because each memory trace contributes additional variance to the match strengths for both target and distracter items, increasing the overlap between these distributions and decreasing discriminability as a consequence (Clark & Gronlund, 1996).

A majority of these models were also able to predict how strengthening items by additional repetitions in the study list produce increases in the hit rates for strengthened items. This is because of the *additional trace assumption*: repetitions of an item produce additional memory traces that add match strength to target items during retrieval. A consequence of this assumption is that repetitions of items have the same effect on the variance of the match strengths as increasing the length of a study list. This is the basis of the *list-strength prediction*, in which increasing the strength of the studied items reduces performance for non-strengthened items due to the interference from the additional traces produced by the repetitions.

The list-strength prediction of the global memory models was tested by Ratcliff, Clark, and Shiffrin (1990) in a series of experiments with all of them resulting in null list-strength effect. Specifically, weak items were not harmed by the strengthening of other items on the study list and strong items did not benefit from the presence of weak items. This was a challenge for virtually all of the global memory models at the time as well as several other memory models, as it was not possible for them to predict a list-length effect while simultaneously predicting a null list-strength effect (see Shiffrin, Ratcliff, & Clark, 1990 for a technical description of how each memory model fails in these predictions).

A potential solution to this problem posited by Shiffrin et al. (1990) was the *differentiation hypothesis*. This proposes a revision to the learning process in which repetitions instead accumulate into a single strong memory trace that is more responsive to its own cue and less responsive to other cues, thus keeping the number of traces constant as strength is increased and preventing interference from increasing. The differentiation mechanism not only predicts a null list-strength effect, but also predicts a lower false alarm rate

(FAR) as the proportion of strong items increases because strong memory traces are less responsive to non-studied items presented at test (Criss, 2006). This prediction is consistent with a decrease in the FAR that has been found as the proportion of strong items on a list is increased in list-strength paradigms (Criss, 2006; Hirshman, 1995; Stretch & Wixted, 1998). The simultaneous increase in hit rate and decrease in FAR as items are strengthened has been referred to as the *strength-based mirror effect*.

While differentiation was first implemented in the SAM model (Shiffrin et al., 1990), since then the null list-strength effect has become a strong constraint on recognition memory models and several models were designed to accommodate this finding. These models include the retrieving effectively from memory model (REM: Shiffrin & Steyvers, 1997), the model of McClelland and Chappell (1998), as well as the Norman and O'Reilly (2003) neural network model, which incorporated differentiation into the medial temporal lobe cortex layer. The most popular of these explanations is likely to be the REM model, as the predictions from REM's differentiation mechanism have been tested and fit to data in a number of experiments (Criss, 2006, 2009, 2010; Starns, White, & Ratcliff, 2010). Differentiation operates in REM by assuming that repetitions or additional storage time provide opportunities to copy additional features from the original study item into its corresponding episodic trace.

What has not been previously explored is what the boundary conditions on this process might be. That is, under which conditions might a repetition of a study item cause a new trace to be created rather than instigating the differentiation of an older trace? To our knowledge, all simulations of differentiation models assume that the first presentation of an item within a study list creates a new memory trace with subsequent presentations differentiate the traces created earlier. Considering these are simulations of experiments where the studied materials were stimuli that were learned prior to the experiment (words), we can expect the study items to have memory traces that were created prior to the experiment. For instance, the word "dog" may have been experienced hours or days prior to the onset of an experiment, producing a pre-experimental memory trace. Without an understanding of the boundary conditions of differentiation, it is not clear why the first presentation of the word "dog" on a study list creates a new episodic trace instead of differentiating the older, pre-experimental episodic trace.

One hypothesis for differentiation's boundary conditions comes from Criss (2009), who stated that new traces are created in new contexts and repetitions within the same context cause differentiation to occur. Thus, if a novel list-strength paradigm was constructed in which repetitions occurred in new contexts and all of the contexts were cued during a test phase, Criss's hypothesis predicts that there should be increased interference as the proportion of repeated items is increased, manifesting in a decrease in discriminability.

In order to test this hypothesis in the novel list-strength paradigm described, there needs to be an understanding of what type of context shift would be sufficient to break the differentiation process. While many episodic memory models discuss the effects of context (Dennis & Humphreys, 2001; Howard & Kahana, 2002; Shiffrin & Steyvers, 1997), context has not yet been operationally defined and instead tends to be described loosely as a set of internal and external elements that enable subjects to focus retrieval on an event, namely the experience of the study list (Klein, Shiffrin, & Criss, 2007). Because there is no theory of context, consideration of a context shift must instead come from empirical work on the subject.

While one might be inclined to think that a shift in environmental context (i.e.: changing the room or other aspects of the physical environment during the study episode) might be sufficient, it has not been shown to produce changes in *d'* in recognition memory (Godden & Baddeley, 1980; Murnane, Phelps, & Malmberg, 1999). One factor that has been shown to make significant contextual shifts is list context (presenting items in different study lists). Attempts at modeling the list discrimination paradigm have measured significant changes in context across different lists (Criss & Shiffrin, 2004; Dennis & Humphreys, 2001). In a recent review of context change manipulations, Klein et al. (2007) suggested that adding testing between list presentations increases contextual separation.

Considering these are the strongest context change manipulations that are currently known to affect recognition memory, we decided to test Criss's (2009) hypothesis by presenting participants with three study-test cycles and repeating items across different study lists rather than within the lists themselves (as is traditionally done in list-strength experiments) and then subsequently testing all of the studied lists with an end-of-session test (this procedure had been used in a list-strength design by Murnane & Shiffrin, 1991, although the repetitions did not occur between list presentations).

If these conditions are sufficient to prevent differentiation, differentiation models predict that a list-strength effect should result and discriminability should decrease as the proportion of strong items increases (see Criss, 2006 for simulations of REM predicting a list-strength effect when differentiation is not used). However, it is also entirely possible that these shifts in context between presentations are *not* sufficient to break the differentiation process. If this is the case, differentiation models predict that there should be a null list-strength effect and a decrease in the FAR as the proportion of strong items increases (Criss, 2006).

In most of the original list-strength experiments, Ratcliff et al. (1990) utilized the *mixed/pure* paradigm. This paradigm uses three conditions: pure strong ("PS": 100% strong items), pure weak ("PW": 100% weak items), and a mixed condition consisting of both strong and weak items (half strong and half weak). A list-strength effect would be

observed if PW > MW (weak items from mixed lists) and if MS (strong items from mixed lists) > PS. In our case, since repetitions are occurring across different lists instead of within a single study list, a pure strong manipulation would entail three repetitions of the same study list. Upon realization that the study lists are the same, it's possible that participants would decrease their attention to the studied items, reducing their performance relative to the mixed condition and artificially inducing a list-strength effect (PS < MS). Instead, both of our experiments used the *mostly strong/mostly weak* manipulation (Ratcliff et al., 1990; Experiment 7), in which all sessions were mixed but the relative proportions of strong and weak items were manipulated. A positive list-strength effect would be observed if performance was worse in the mostly strong condition than in the mostly weak condition.

Experiment 1

Method

Participants A total of 108 participants contributed to this study. Participants were undergraduate students participating for course credit.

Stimuli The word pool consisted of 264 words that were between five and seven letters in length and between 1 and 4 Google counts per million in normative word frequency (low frequency).

Procedure Participants were randomly assigned to either the mostly strong or mostly weak conditions. In both conditions, 180 items were presented across three study/test cycles. For the mostly strong condition, 80% of the items (144 items) were strengthened by presenting them in every study list and the other 20% (36 items) were distributed across the study lists (12 in each list), leading to a total of 156 presentations for each study list. For the mostly weak condition, 80% of the items were distributed across the three study lists (48 in each list) and 20% of the items were strengthened by presenting them in every study list, leading to a total of 84 presentations per list.

Before the study phase, participants were instructed to study the items and make a pleasantness rating from 1-4 on the keyboard while viewing the word. Each word was studied for 2.5 seconds and followed by a blank screen for 500 ms.

Following each study list, participants engaged in a distracter task consisting of a card game in which cards from a deck were presented one at a time and participants had to hit the spacebar when different patterns of cards had been presented (such as two cards of the same suit in a row). Participants engaged in distracter activity for 180 seconds in the mostly strong condition and for 396 seconds in the mostly weak condition. The purpose of the different lengths of the distracter task was to ensure that the time between the start of the study list and the beginning of the test list was equivalent in both conditions. After the three study/test cycles had completed, participants engaged in another 240 seconds of distracter activity before the end-of-session test.

Before each test list prior to the end-of-session test, participants were instructed to say "yes" to items they had studied during the preceding study list and "no" otherwise. These test lists consisted of 20 test items. In the mostly strong condition, 7 of the test items were drawn from the strong items and 3 of the items were drawn from the weak items. In the mostly weak condition, 7 of the items were drawn from the weak items and 3 of the items were drawn from the strong items. The remaining 10 test items were distracters that had not been presented on any of the study lists. The purpose of using different numbers of test items in the different conditions was to ensure that there would be enough test items in the end-of-session test for the minority group of each condition. No test items were repeated across the different test lists. The test lists were self paced and a blank screen followed each presentation for 500 ms.

After the three study-test cycles and the distracter activity had completed, participants began the end-of-session test. Participants were instructed that they were going to be tested on all of the items they had previously studied and that they were to respond "yes" to any items they had studied during the session and "no" otherwise. Prior to this instruction, participants were given no indication that they would be re-tested on the study lists in this fashion. Participants were tested on 27 strong items, 27 weak items, and 54 distracter items. An equal number of weak items were sampled from each of the studied lists and none of the end-of-session test items had been previously tested.

Results and Discussion

d' scores were calculated for all participants as a measure of memory sensitivity for the end-of-session test. To avoid infinite values of *d'*, edge corrections were performed on the hit and false alarm rates by adding 0.5 to the hit and false alarm counts and 1 to the target and distracter counts (Snodgrass & Corwin, 1988).

Data from 8 participants were not used in the analysis due to failure to finish the experiment. All participants who finished the experiment exhibited above chance performance ($d' > 0$). Hit rates, false alarm rates, and *d'* scores from the end-of-session test can be seen in Figure 1. It should be noted that while the graphs separate the hit rates for weak items into their study list of origin, all statistical analyses collapsed across the hit rates.

Because our hypotheses are concerned with the presence of a list-strength effect when multiple lists are cued, data analysis was restricted to performance from the end-of-session test. Results indicated that strong items were better discriminated than weak items, $F(1, 98) = 271.20$, $p < .001$. Because both list conditions are mixed, a list-strength effect would be observed if performance in the mostly weak condition exceeded that of the mostly strong condition. However, a null list-strength effect was observed:

Performance on the mostly weak condition did not exceed that of the mostly strong condition, $F(1, 98) = 1.47$, $p > .05$.

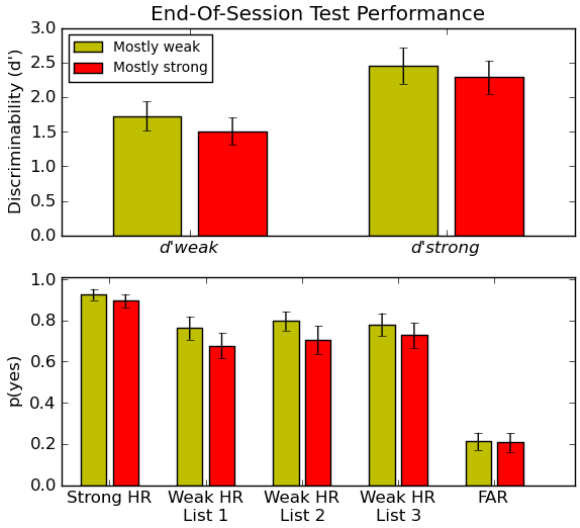

Figure 1: Performance from the end-of-session test of Experiment 1. Top – *d'* scores for weak and strong items in each condition. Bottom – false alarm rates and hit rates for strong and weak items. The hit rates for weak items are separated by the list they were originally studied in. Error bars represent the 95% confidence interval.

One might be inclined to think that the context shifts that occurred between repetitions of items weren't dramatic enough to break the boundaries of differentiation. However, the decrease in FAR that differentiation models predict as the proportion of strong items increases was not found, $t(97.27) = 0.14$. Thus, these results present a challenge for differentiation models.

While the results of the data analysis suggest a null list-strength effect, there was a trend in the direction of a list-strength effect, as hit rates for weak items were worse in the mostly strong condition than in the mostly weak condition, $t(90.47) = 2.22$, $p < .05$, although hit rates for strong items were not significantly worse in the mostly strong condition, $t(94.47) = 1.43$, $p > .05$. One possible reason why this might be is because the lengths of the lists in the mostly strong condition (156 presentations per list, which is almost eight minutes) are unusually long for recognition experiments. Many participants seemed very frustrated with the length and tedium of the task, which may have caused fatigue. Based on these observations, we decided to run the same task with a smaller number of items to see if we would obtain the same results.

Experiment 2

The design of Experiment 2 was identical to that of Experiment 1 except with a reduction in the total number of items studied in an attempt to reduce the possible fatigue that may have resulted in the trend found in Experiment 1.

Method

Participants A total of 104 participants contributed to this study. Participants were undergraduate students participating for course credit.

Procedure The procedure was identical to that of Experiment 1 except that the total number of presented items was 120 instead of 180. Thus, in the mostly strong condition, 96 items were strengthened by presenting them in all three lists and 24 items were distributed across the three lists (eight items per list), totaling 104 presentations per list. In the mostly weak condition, 96 items were distributed across the three lists (32 items per list) and 24 items were strengthened by presenting them in each list, totaling 56 presentations per list. Following each study list, distracter activity commenced for 180 seconds for the mostly strong condition and 324 seconds for the mostly weak condition. The end-of-session test consisted of 15 strong items, 15 weak items, and 30 distracter items.

Results

All participants exhibited above chance performance ($d' > 0$). Hits, false alarm rates, and *d'* scores for the end-of-session test data can be seen in Figure 2.

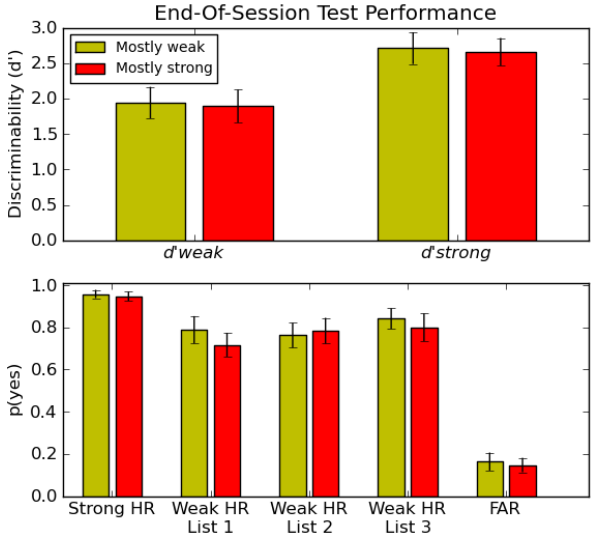

Figure 2: Performance from the end-of-session test of Experiment 2. Top - *d'* scores for strong and weak items in each condition. Bottom – false alarm rates and hit rates for strong and weak items in each condition. The hit rates for weak items are separated by the list they were originally studied in. Error bars represent the 95% confidence interval.

Results from Experiment 2 replicated the key results of Experiment 1. A null list-strength effect was observed in that discriminability did not change across the two conditions, $F(1, 102) = 0.11$, strong items were better discriminated than weak items, $F(1, 102) = 272.76$, $p < .001$, and FAR did not differ between the two conditions,

$t(97.96) = .59$. Unlike in Experiment 1, hit rates did not differ between the two conditions for weak items, $t(94.57) = 1.04$, $p > .05$, or strong items, $t(100.56) = 0.72$.

General Discussion

The differentiation hypothesis has become a popular explanation for the null list-strength effect in recognition memory. While the predictions of differentiation have been tested, little work has been done to explore or test the boundary conditions under which such a process might occur, despite the necessity in specifying such conditions in order to understand why a memory trace would be created upon an item's first presentation in a study list and differentiated in subsequent presentations.

Criss (2009) proposed the hypothesis that new traces are created in new contexts. We tested this hypothesis by using the strongest context shifts that have been reported to affect recognition memory by repeating items across different study/test cycles rather than within a single study list as is traditionally done in list-strength experiments. When the different lists were all cued in a final end-of-session test, increasing the proportion of strong items that were studied did not produce a list-strength effect and also did not produce a significant decrease in the false alarm rate.

These results present a challenge for differentiation models, which yield predictions that cannot account for both of these null effects. If differentiation was not occurring while viewing the study lists, differentiation models would predict a list-strength effect due to the increased interference from the creation of new memory traces during repetitions. If differentiation was not occurring due to the context shift not being dramatic enough, differentiation models would predict a decrease in the false alarm rate due to the strong memory traces providing a weaker match to the distracter items.

It should be mentioned that differentiation is not required to explain the data from list-strength paradigms. Context-noise models, such as the bind-cue-decide model of episodic memory (BCDMEM: Dennis & Humphreys, 2001) and the recognition model proposed by Anderson and Bower (1972), predict a null list-strength effect because the recognition decision is based on the match between the test context and an item's retrieved contexts. Because non-probe items do not affect the decision, the strength of those items cannot provide interference.

The decrease in the FAR that tends to occur in list-strength paradigms can also be explained by a criterion shift. This account states that as the proportion of strong items on a study list increases, participants are more likely to use a stricter criterion at test that reflects higher expectations of their learning from the study episode (Hirshman, 1995; Starns et al., 2010; Stretch & Wixted, 1998). This account was directly tested by Starns et al. (2010) in a series of experiments in which they found that retrieval expectations at test, not encoding conditions, produced changes in the FAR. This was tested by conducting a list-strength experiment in which participants were informed of the strength composition of the test list before being tested. The FAR decreased when participants were told that only the strong items would be tested and increased when they were told that only the weak items would be tested, despite the fact that encoding conditions were identical under both sets of instructions.

Acknowledgements

We would like to thank members of the Memory and Language Lab (MALL) and the Computational Memory Lab (CML) at The Ohio-State University for their suggestions and support.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*, 97–123.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*, 37-60.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*, 461–478.

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology*, *59*, 297-319.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 484-499.

Criss, A. H. & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review*, *111*(3), 800-807.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*, 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. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

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

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5–16.

Godden, D. R., & Baddeley, A. D. (1980). When does context influence recognition memory? *British Journal of Psychology, 71*, 99-104.

Hintzman, D. L. (1988). Judgments of frequency and

recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528-551.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 345-351.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269-299.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 171-189). New York: Psychology Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 734–760.

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

Murnane, K., & Shiffrin, R. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855-874.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General*, *128*, 403 415.

Norman, K. A. & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*, 611-646.

Pike, R. (1984). A comparison of convolution and matrix distributed memory systems. *Psychological Review, 91,* 281-294.

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

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

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

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language*, *63*, 18-34.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379-1396.