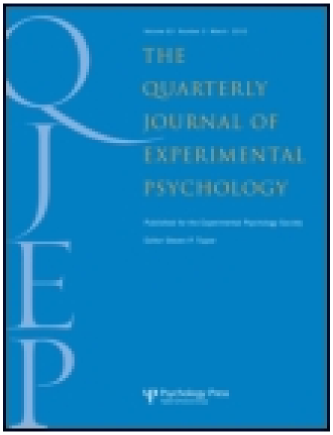

The Quarterly Journal of Experimental Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

Stimulus type and the list strength paradigm

Adam F. Osth[a], Simon Dennis[b] & Angela Kinnell[c]

[a] Department of Psychology, The Ohio-State University, Columbus, Ohio, USA

[b] School of Psychology, University of Newcastle, Australia

[c] School of Psychology, University of Adelaide, Australia

Published online: 03 Mar 2014.

To cite this article: Adam F. Osth, Simon Dennis & Angela Kinnell (2014) Stimulus type and the list strength paradigm, The Quarterly Journal of Experimental Psychology, 67:9, 1826-1841, DOI: 10.1080/17470218.2013.872824

To link to this article: http://dx.doi.org/10.1080/17470218.2013.872824

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2014
Vol. 67, No. 9, 1826–1841, http://dx.doi.org/10.1080/17470218.2013.872824

Stimulus type and the list strength paradigm

Adam F. Osth[1], Simon Dennis[2], and Angela Kinnell[3]

[1]Department of Psychology, The Ohio-State University, Columbus, Ohio, USA
[2]School of Psychology, University of Newcastle, Australia
[3]School of Psychology, University of Adelaide, Australia

In recognition memory, increasing the strength of studied items does not reduce performance on other items, an effect dubbed the null list strength effect (LSE). While this finding has been replicated many times, it has rarely been tested using stimuli other than single words. Kinnell and Dennis (2012) recently tested for the presence of list length effects using non-word stimulus classes while controlling for the confounds that are present in list length designs. Small list length effects were found for fractal and face images. We adopted the same paradigm and stimuli used by Kinnell and Dennis to test whether these stimuli would be susceptible to list strength effects as well. We found significant LSEs for fractal images, but null LSEs for face images and natural scene photographs. Stimuli other than words do appear to be susceptible to list strength effects, but these effects are small and restricted to particular stimulus classes, as is the case in list length designs. Models of memory may be able to address differences between these stimulus classes by attributing differences in representational overlap between the stimulus classes.

Keywords: Recognition memory; Global matching models; Computational modelling; Criterion shifts; List strength.

A central issue in recognition memory research is specifying the source of interference. A class of memory models referred to as the *global matching models* characterized the studied items as the source of interference in recognition memory; this has been referred to as the *item noise* conception of interference. Global matching models posit that retrieval in recognition memory is a parallel search of the contents of memory in response to a probe cue, producing a "familiarity" value that indexes the similarity of the probe item to the contents of memory (Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989). A success of the global matching models is the ability to account for the *list length effect*, which is a decrease in performance as the number of items on a study list is increased (Strong, 1912). List length effects are predicted because each item that is added to memory spuriously overlaps with a probe cue. The mean and variance of the resulting familiarity distributions for targets and lures are equivalent to the sum of the means and variances of the matching strengths between the probes and each item in memory. Thus, each item that is added to the contents of memory increases the variance of the resulting familiarity distributions, which causes increased overlap between the distributions for targets and distracters and lower discriminability as a consequence (Clark & Gronlund, 1996).

Correspondence should be addressed to Adam Osth, 1835 Neil Avenue, Columbus, OH 43201. E-mail: adamosth@gmail.com
We would like to thank William Hockley and two anonymous reviewers for their helpful comments on the original version of this manuscript.

What is not immediately obvious about the global matching models is that the *strength* of the studied items influences the amount of interference. If an item is added into memory multiple times, there are multiple memory components corresponding to the studied item that all spuriously overlap with a probe cue. Thus, repetitions have the same functional effect as increasing the length of a study list[1] and the global matching models predict that discriminability should decrease as the strength of other items or the proportion of strengthened items (these two variables are collectively referred to as *list strength*) is increased.

The list strength prediction of the global matching models was tested in a series of experiments by Ratcliff, Clark, and Shiffrin (1990) by varying the list composition of strengthened vs. non-strengthened items; items were strengthened by both study time and repetition. All seven of their experiments revealed that the strength of the studied items had no impact on performance: weak items were not harmed when other items on a study list were strengthened, and strong items did not benefit by being accompanied by weak items instead of strong items on the study list. The *null list strength effect* in recognition memory has been replicated several times since (Hirshman, 1995; Kahana, Rizzuto, & Schneider, 2005; Murnane & Shiffrin, 1991a, 1991b; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Shiffrin, Huber, & Marinelli, 1995; Yonelinas, Hockley, & Murdock, 1992), although all of these investigations have used words or sentences as stimuli. The null list strength effect was one of the findings that led to the abandonment of the first wave of global matching models and inspired the development of newer models such as the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997), the Subjective Likelihood in Memory model (SLiM; McClelland & Chappell, 1998) and the Bind Cue Decide Model of Episodic Memory (BCDMEM; Dennis & Humphreys, 2001).

Dennis and Humphreys (2001) argued that null effects of list strength can be understood within a global matching framework if the representations of word stimuli are sufficiently sparse such that they do not overlap with each other. If the overlap between word representations is negligible, the probe cue only matches its own stored item representation in memory and bears no match to the other stored item representations. Consequently, both the number and strength of other stored item representations does not cause interference at retrieval.

While minimal representational overlap can predict a null list strength effect, it also predicts a null list length effect in recognition memory performance. Despite earlier findings of a positive list length effect, Dennis and Humphreys (2001) noted a number of confounds in list length designs that have artefactually contributed to the finding of a positive list length effect. For instance, when immediate testing is used in a list length design, the retention interval is longer for the early items on a long list compared to a short list. When such confounds are controlled, all of the experiments by Dennis and colleagues have found no effect of list length on recognition memory performance when single words are used as stimuli (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011) or when word pairs are used in an associative recognition task (Kinnell & Dennis, 2012).

The theory proposed by Dennis and Humphreys (2001) argues that interference does not come from the list items but instead comes from the prior contexts in which an item has been experienced; this notion has been referred to as *context noise*. According to context noise models, the greater the number of exposures in prior contexts, the greater the interference at retrieval.

List length effects and stimulus interactions

The inspiration for the present experiments came from the list length experiments conducted by

[1]Technically, list strength effects were predicted to be larger in several global matching models because multiple memory components corresponding to a studied item all have covariance in their match to a probe cue. This additional covariance produces a larger decrement in performance than would be predicted for a manipulation of list length.

Kinnell and Dennis (2012) using non-word stimuli. While a number of experiments using word stimuli have found null effects of list length on recognition-performance, an investigation by Kinnell and Dennis (2012) found positive effects for specific non-word stimulus classes, namely novel fractal images and novel faces, while null effects of list length were found for natural scene photographs.

Why would novel non-word stimuli be susceptible to list length effects? Dennis and Humphreys argued that a pure context noise theory can only be sensibly applied to stimuli that have sufficient background experience, as a stimulus that has never been seen before has no association to pre-experimental contexts and therefore cannot suffer from context noise at retrieval. Thus, given that novel non-word stimuli cannot be subject to context noise, it seems reasonable to assume that these stimuli suffer from item noise at retrieval. A global matching model is capable of predicting item noise effects if the item representations for a given stimulus class overlap with each other, making them susceptible to effects of list length and list strength at retrieval.

These experiments investigate the latter prediction from this account. Novel non-word stimuli are not necessarily guaranteed to suffer interference from a list strength manipulation. Differentiation models of episodic memory, for instance, predict a positive list length effect and a null list strength effect (Shiffrin, Ratcliff, & Clark, 1990, more on differentiation models is discussed in the General Discussion). In our investigation, we used a paradigm very similar to the list length experiments by Kinnell and Dennis (2012) and tested fractal images, face images, and natural scene photographs using a list strength paradigm. Participants experienced two list strength conditions: a pure weak list where each item was only presented once and a mixed list where half the items were presented once and the other half were presented four times. A positive list strength effect would be observed if performance on once-presented items is worse in the mixed list relative to the pure weak list. To make the results comparable to those of Kinnell and Dennis, we employed the controls used in the list length experiments by Dennis and colleagues (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011, 2012), as the confounds in list length designs can artefactuallycontribute to the finding of a list strength effect (more on this is discussed in the General Method section).

Preliminary findings indicated that there were large criterion shifts in response to the list strength manipulation. A weakness of using the standard equal variance signal detection model is that when a criterion shift is taking place, d' and c becomecon-founded with each other. To ensure that our results weren't restricted to the choice of signal detection model employed, we additionally collected data using a two alternative forced choice (2AFC) testing procedure.

To clarify exposition of the experiments within, we begin with a General Method section that describes the methodology for each experiment, as each of the three experiments had only minor deviations from the general method. Explication of more specific details, such as the type of encoding task and the length of the distracter activity, can be found in the Method sections for the individual experiments. We then discuss the list strength experiments conducted on fractal images, face images, and natural scene photographs. To foreshadow our results, we found a positive list strength effect with fractal images in both the yes/no and 2AFC testing while null effects of list strength were found for face images and natural scene photographs in both the yes/no and 2AFC tests.

GENERAL METHOD

Participants

All participants were undergraduate students from The Ohio-State University who participated for course credit in an introductory psychology course.

Procedure

A diagram of the basic procedure can be seen in Figure 1. Each participant participated in two conditions: a pure weak list composed entirely of

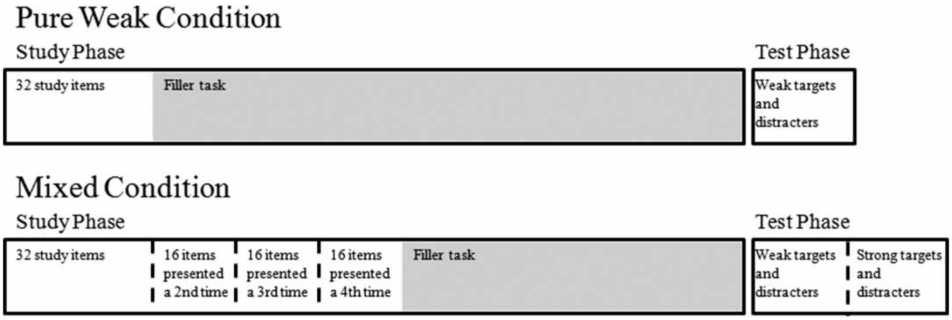

Figure 1. *Graphical depiction of the general method employed in all three experiments. In the pure weak condition, all items are studied once, participants engage in a long filler task, and are then tested on targets and distracters. In the mixed list condition, all items are studied once, and then half of the items are repeated an additional three times. Participants then engaged in a shorter filler task and in the test phase they were tested on all of the weak targets and distracters before being tested on an additional block of strong targets and distracters. Memory was tested using either yes/no or two alternative forced choice (2AFC) tests.*

once-presented items as well as a mixed list composed of the same number of items, but half the items were presented once and half were presented four times. The order of the conditions was counterbalanced across participants. Type of testing (yes/no vs. 2AFC) was manipulated between subjects. To ensure that participants paid sufficient attention during the study list presentations, participants engaged in a rating task in two of the three experiments (in Experiment 3, which used natural scene photographs, the presentation time was too rapid for a rating task to be employed at study).

In the mixed list condition, all items were presented once and, subsequently, half the items were repeated an additional three times. During the repetitions, all the strong items were shuffled and presented before an additional repetition occurred. That is, if a given strong item was presented twice then all other strong items were presented twice before a third presentation occurred. After the completion of the study list, participants completed a distracter task in which images of playing cards appearedon the screen and participants were instructed to press the space bar when they observed specific sequences (such as two cards in a row that shared the same suit, or two cards that summed to 11). The distracter task took place for at least 180 seconds, although the duration was longer in the pure weak list to ensure that the time between the beginning of the study list and the beginning of the test list was the same for both conditions. The specific length of the distracter task for the pure weak list depended on the presentation time of the items, which varied across the experiments, and therefore can be found in the specific Method sections for each experiment.

In the yes/no recognition test, participants were instructed to respond "yes" to all studied items and "no" otherwise. In the 2AFC test, participants were presented with both a studied and an unstudied item on either the left- or right-hand sides of the screen and participants were instructed to select the item they recognized from the study list. In half of the 2AFC trials the target item was on the left-hand side of the screen and in the other half it was on the right-hand side of the screen. Response keys were the same for both the yes/no and 2AFC tests (1 and 0).

In the mixed condition there were two blocks of test items: a weak block of items that consisted of weak targets and unstudied items, and a subsequent strong block of items that consisted of strong targets and unstudied items. The strong items were tested after the weak items because it has been found that performance decreases during recognition memory testing (Criss, Malmberg, & Shiffrin, 2011; Ratcliff & Murdock, 1976), and a mixed block of strong and weak items would make it such that weak items would on average be tested later in the mixed list than in the pure weak list, decreasing performance and artefactually contributing to a list strength effect. It was for this reason that we only employed a pure weak and mixed list: the inclusion of an additional "pure strong" condition with entirely strong items would make it impossible to simultaneously control for the differences in retention interval and test position across all three conditions.

All experiments were designed and run using the Python experimental library (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007).

Analysis

For simplicity, performance in the yes/no test of each experiment was measured using d' calculated according to equal variance signal detection measures. However, one problem with using such a measure is that d' and c are confounded with each other when the memory strength distributions for studied and unstudied items have unequal variance (see Rotello, Masson, & Verde, 2008, for a simulation study and description of how the wrong choice of signal detection model can inflate Type I error rates). Specifically, decreases in the hit rate produce larger decreases in d' as the variability of the target distribution is increased. In item recognition, it has generally been found that targets have higher variability than distracters, with targets having a standard deviation approximately 1.25 times larger (Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff et al., 1992, 1994). However, it should be mentioned that these investigations have used words as stimuli and the results are not guaranteed to generalize to the novel non-word stimulus classes that we have employed.

While it would be possible to collect confidence ratings and calculate d_A as a non-parametric measure of performance, an unsolved problem for signal detection models is that changes in bias distort both the slope and intercept of the z-transformed receiver operating characteristic (Van Zandt, 2000). Given that there were large shifts in bias in our experiments, usage of d_A may not be appropriate. Instead, we rely on proportion of correctly chosen targets ($p(c)$) in the 2AFC test conditions as a non-parametric measure of performance.

To avoid infinite values of d', edge corrections were performed by adding .5 to the hit and false alarm counts and 1 to the target and distracter counts (Snodgrass & Corwin, 1988). This correction was only performed for the signal detection analyses, and all statistical analyses on response rates were on the raw, uncorrected response rates.

EXPERIMENT 1

Experiment 1 used fractal images as stimuli. A fractal is a geometric shape in which each component is a smaller representation of the whole and can be produced iteratively via recursion. Kinnell and Dennis (2012) argued that fractal images have ahigher degree of intra-stimulus perceptual similarity due to the lack of a common structure and the difficulty in applying labels to segments of the image. They found significant effects of list length in response latency and conjectured that fractal images may be subject to this form of item noise due to the overlap between their representations. We conducted a list strength experiment with fractal images to evaluate whether the strength of the fractal images produces interference.

Method

Participants

A total of 200 participants contributed to this experiment (100 for the yes/no test).

Materials

The stimuli in all experiments were identical to those used by Kinnell and Dennis (2012). In this experiment, stimuli were 140 fractal images that were 800×600 pixels in size. Half of the fractals were based on circles and the other half had leaf-like shapes. Example fractal images can be seen in Figure 2. In the 2AFC tests, the fractal images were reduced to 600×450 pixels in size such that two images could fit on the screen side by side.

Procedure

The design of this experiment is described in the General Method.

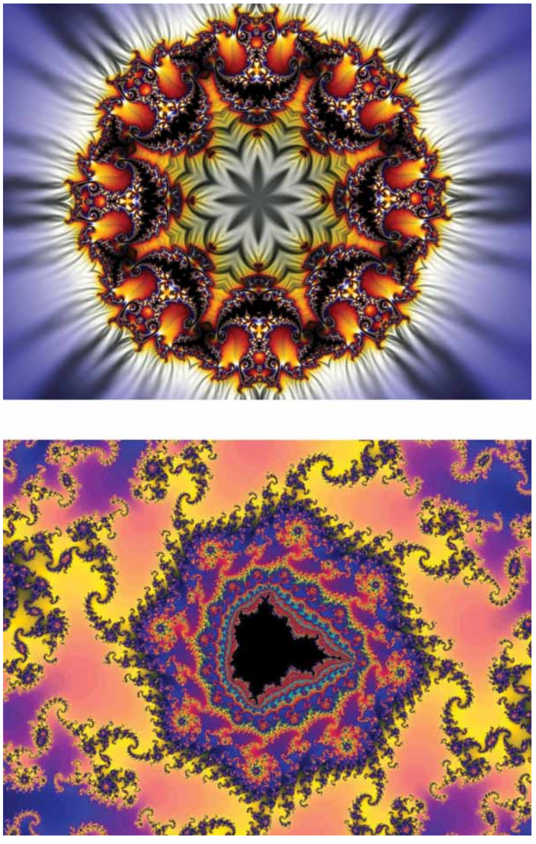

Figure 2. *Examples of a circle (top) and leaf (bottom) fractal image used in Experiment 1.*

A total of 32 items were presented to the participants in each study list at 3000 ms per item with a blank screen occurring between presentations for 250 ms. Participants were instructed to rate the pleasantness of each item during the study phase on a four-point scale and enter the rating using a key on the keyboard. To ensure that the retention intervals were equivalent across the two conditions, the length of the distracter activity after the pure weak list was 366 seconds.

Results

Performance measures such as d', c, and $p(c)$ for all experiments can be seen in Figure 3, while response rates can be seen in Figure 4. Data from twenty participants were excluded for having at or below chance

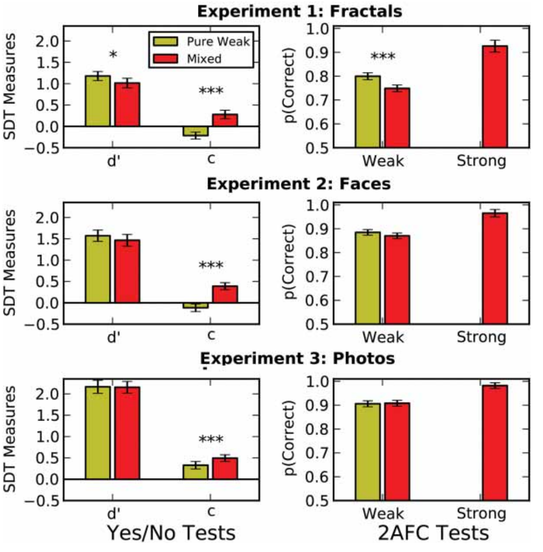

Figure 3. *Mean performance measures from all experiments, including signal detection measures d′ and c for the yes/no tests and probability of correct choice (p(c)) for the 2AFC tests. Error bars represent 95% within-subjects confidence intervals.*

performance ($d' <= 0$ in the yes/no tests, $p(c) <= .5$ in the 2AFC tests), twelve were from the yes/no test condition.

Separate one-way analyses of variance (ANOVAs) were calculated for each dependent variable of interest in the yes/no and 2AFC tests. A positive list strength effect was found, in that performance was significantly worse in the mixed list relative to the pure weak list in both the yes/no tests, $F(1, 87) = 4.96$, $p < .05$, $\eta_p^2 = .05$, and in the 2AFC tests, $F(1, 91) = 12.42$, $p < .001$, $\eta_p^2 = .12$.

In the yes/no tests, responding was significantly more conservative in the mixed list relative to the pure weak list, as reflected by higher values of c, $F(1, 87) = 98.77$, $\eta_p^2 = .53$, lower hit rates, $F(1, 87) = 92.09$, $\eta_p^2 = .51$, and lower false alarm rates, $F(1, 87) = 59.80$, $\eta_p^2 = .40$, all ps $< .001$. While both the hit and false alarm rates were lower in the mixed list, the decrease in the hit rate is larger, which is the locus of the list strength effect.

In the mixed list, strong hit rates were higher than weak hit rates in the yes/no tests,

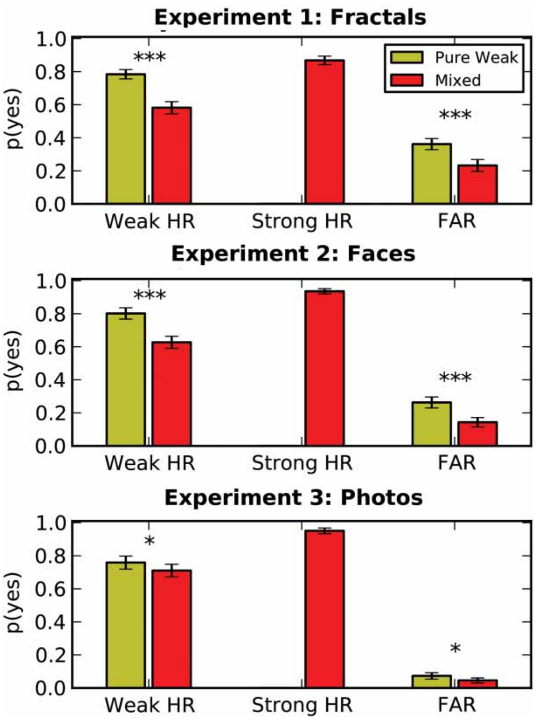

Figure 4. *Mean response rates for the yes/no tests from all experiments, including hit rates (HR) and false alarm rates (FAR). FAR for the mixed list conditions are the FAR from the weak block. Error bars represent 95% within-subjects confidence intervals.*

$p(c)(87) = 9.59$, and $p(c)$ for strong items was higher than for weak items in the 2AFC tests, $t(91) = 12.57$, both $ps < .001$. In the yes/no test, there were no significant differences between false alarm rates in the weak (.23) and strong (.22) blocks of items, $t < 1$.

Discussion

We found a significant positive effect of list strength on recognition memory performance when fractal images are used as stimuli, in that performance was significantly worse in the mixed list relative to the pure weak list in both the yes/no and 2AFC tests. Results of this experiment were consistent with the results of Kinnell and Dennis (2012), who found a significant list length effect on recognition memory performance when fractal images are used as stimuli. Taken together, these results indicate that fractal images are generally susceptible to the item noise effects predicted by the first wave of global matching models.

While the effect size was lower in the yes/no tests than in the 2AFC test ($\eta_p^2 = .05$ and .12, respectively), this may be due to the equal variance signal detection measures employed in the signal detection analysis of the yes/no test data. When d_A is substituted for d' with a target variability that is 1.25 times larger than the variability for distracters, the differences in performance between the two conditions becomes larger and the effect size becomes comparable to that of the 2AFC test ($\eta_p^2 = .128$). Nonetheless, this analysis is purely speculative and a proper evaluation of the ratio of standard deviations for fractal images is required.[2]

EXPERIMENT 2

In Experiment 2, images of novel faces were used as stimuli. Kinnell and Dennis (2012) found small but significant effects of list length on discriminability for faces and argued that while faces conform to a familiar structure, novel faces are themselves unfamiliar and contain many idiosyncratic features, which may possibly lead to overlap in their representations. Converging evidence from this hypothesis comes from Norman, Tepe, Nyhus, and Curran (2008), who found a significant but small effect of list strength on discriminability when faces were used as stimuli.

Method

A total of 201 participants contributed to this experiment (105 for the yes/no test).

[2]While we have argued that interpretation of receiver operating characteristic (ROC) measures is hindered by the distortions in slope and intercept that occur with shifts in bias (Van Zandt, 2000), an alternative strategy is to measure the variability of drift rates using response time models (Starns, Ratcliff, & McKoon, 2012). However, this approach tends to require a much larger number of observations for both correct and error responses than we have collected in these experiments.

Materials

The stimuli were 140 colour images of faces from the AR Face Database (Martinez & Benavente, 1998). Half of all the faces were male faces and the other half were female faces. Stimuli that were targets on the recognition memory test were identical to the images that were studied on the study list. All images were 460×460 pixels in size. Examples of possible face stimuli can be seen in Figure 5.

Procedure

The procedure of this experiment was identical to the one used in Experiment 1. All participants saw an equal number of male and female faces as weak, strong, and distracter items.

Results

Data from ten participants were excluded for having at or below chance performance (six were from the yes/no test). Data from an additional four participants were excluded for failure to finish the experiment (three were from the yes/no test).

There was no significant effect of list strength on recognition memory performance, in that performance was not significantly worse in the mixed list relative to the pure weak list in either the yes/no test, $F(1, 95) = 1.78$, $\eta_p^2 = .0002$, or in the 2AFC test, $F(1, 90) = 1.46$, $\eta_p^2 = .016$, both ps $> .05$. Like in Experiment 1, responding in the yes/no test was significantly more conservative in the mixed list relative to the pure weak list, as the mixed list exhibited higher values of c, $F(1, 95) = 121.27$, $\eta_p^2 = .56$, lower hit rates, $F(1, 95) = 64.51$, $\eta_p^2 = .40$, and lower false alarm rates, $F(1, 95) = 59.89$, $\eta_p^2 = .38$, all ps $< .001$.

In the mixed list, strong hit rates were higher than weak hit rates in the yes/no test, $t(95) = 16.11$, and $p(c)$ for strong items was higher than for weak items in the 2AFC test, $t(90) = 8.43$, both ps $< .001$. In the yes/no test, there was a marginally significant difference between the false alarm rates in the weak (.14) and strong (.12) blocks, $t(95) = 1.87$, $p = .06$, although the difference is quite small.

Discussion

The results of this experiment contrasted with those of Experiment 1 in that there was no effect of list strength on recognition memory performance in either the yes/no test or the 2AFC test. However, like in Experiment 1, much more conservative responding was exhibited in the mixed list

Figure 5. *Examples of a male and female face image from the AR Face Database.*

relative to the pure weak list, with much lower hit and false alarm rates in the mixed list condition of the yes/no test conditions.

It was rather surprising that we found a null list strength effect given that Kinnell and Dennis (2012) found a positive list length effect using the same stimuli in their experiments. One possible reason for this discrepancy is that repetitions might be encoded with less efficacy than presentations of novel stimuli due to their redundance. Several memory models employ an encoding mechanism in which redundant stimuli are encoded less effectively than novel stimuli (Farrell & Lewandowsky, 2002; Lewandowsky & Murdock, 1989; McDowd & Murdock, 1986; Murdock, 2003); such a mechanism produces weaker learning for repetitions than for novel items and would predict less interference in a list strength paradigm than would be seen in a list length paradigm. We would also like to emphasize that the list length effect found in the Kinnell and Dennis study was rather small in magnitude. In our experiment, both the d' and $p(c)$ were lower in the mixed list than in the pure weak lists, but the differences were far too small to be detected.

It was also somewhat surprising that we did not replicate the positive list strength effect found using novel faces by Norman et al. (2008). One possible explanation for this difference is that the face images used in their experiments were artificially generated and lacked many of the idiosyncratic features that are present in the novel faces used in our experiments. It seems likely that their faces were more confusable with each other, potentially leading to more overlapping representations and suffering from more item noise as a result.

EXPERIMENT 3

In Experiment 3, photographs of natural scenes were used as stimuli. Unlike with fractal and face images, Kinnell and Dennis (2012) found no significant effects of list length on performance for photographs of natural scenes. They argued that this result may be because natural scene photographs are somewhat "wordlike." That is, natural scene photographs tend to be easily labelled and can be encoded using semantic features in a much easier fashion than would be possible for other pictorial stimuli (such as fractal images). For instance, a participant glancing at the example photos in Figure 6 could easily describe the top and bottom images as "snow-covered mountain" and "beach", respectively.

The idea that natural scene photographs have linguistic representations dates back to Paivio (1971, 1976), who posited dual codes for pictorial representations that contained both perceptual and linguistic information. Evidence for this hypothesis comes from studies showing that recognition performance for pictures is still superior to single words even when the test stimuli are verbal

Figure 6. *Examples of natural scene photographs used in Experiment 3.*

labels of the pictures that were seen on the study list (Paivio, 1976; Madigan, 1983). Similarly, both pictures and words rated high in familiarity exhibit lower hit rates and higher false alarm rates on recognition memory tests than pictures and words rated low in familiarity (a mirror effect; Glanzer & Adams, 1985), with the reverse being true on tests of free recall (better performance for pictures and words rated as familiar; Karlsen & Snodgrass, 2004). If representations of natural scene photographs are indeed like the representations for words, then there should also be no effects of list strength on their recognition performance.

Method

Participants

A total of 162 participants contributed to this experiment (74 for the yes/no tests).

Materials

The stimuli used in this experiment were identical to those used by Kinnell and Dennis's (2012) fourth experiment. The stimuli were 140 colour photographs of everyday scenes. Example photographs can be seen in Figure 6. Each image was 800×600 pixels in size. In the 2AFC tests, the photos were reduced to 600×450 pixels in size such that two images could fit on the screen side by side.

Procedure

The procedure of this experiment was nearly identical to that of Experiment 2, but with a couple of small exceptions. The first is that the presentation time for each stimulus was reduced to 500 ms. This change is largely to offset the high performance that is typically found for recognition of natural scene photographs, as participants' performance levels can be at ceiling if they are given ample study time. Due to the shorter study time, participants were not asked to make pleasantness ratings on the stimuli. Participants were also warned of the short study times and were told to maintain careful attention during the presentation of the study list.

Changing the study time also changed the overall time taken to study the list, so the filler time length for the pure weak list was changed to 300 seconds (the mixed list filler time remained at 270 seconds) to ensure that the time between the start of the study list and the start of the test list was the same across both list strength conditions.

Results

Data from four participants were excluded for having at or below chance performance (three were from the yes/no test condition).

There was no significant effect of list strength on recognition memory performance, in that performance in the mixed list was not significantly worse than the pure weak list in either the yes/no test, $F(1, 70) = .011$, $\eta_p^2 = .0001$, or in the 2AFC test conditions, $F(1, 86) = .03$, $\eta_p^2 = .0003$, both ps $> .05$. In the yes/no tests, responding was significantly more conservative in the mixed list condition relative to the pure weak condition, as the mixed list condition exhibited lower values of c, $F(1, 70) = 13.50$, $\eta_p^2 = .16$, $p < .001$, lower hit rates, $F(1, 70) = 5.30$, $\eta_p^2 = .07$, $p < .05$, and lower false alarm rates, $F(1, 70)$, $= 6.01$, $\eta_p^2 = .08$, $p < .05$.

In the mixed list, strong hit rates were higher than weak hit rates in the yes/no tests, $t(70) = 13.92$, and $p(c)$ for strong items was higher than for weak items in the 2AFC tests, $t(86) = 6.85$, both ps $< .001$. In the yes/no test, the difference between false alarm rates in the weak (.046) and strong (.047) blocks of the mixed list was not significant, $t(70) = -.18$, $p > .05$.

Discussion

Performance for natural scene photographs were similar to those for words in that they exhibited no effect of list strength on recognition memory performance in both yes/no testing and 2AFC testing. Taken together with the null list length effect found for natural scene photographs found by Kinnell and Dennis (2012), these results indicate that performance on natural scene photographs is qualitatively similar to performance for words (aside from the generally superior performance for photographs; see Shepard, 1967; Standing, 1973). Based on these similarities, global matching models may be able to account for performance on pictures and words by assuming that both stimulus classes have unitized item representations.

GENERAL DISCUSSION

The vast majority of recognition memory experiments use only words as stimuli. While many regularities in performance have been discovered across these experiments, these effects are not guaranteed to generalize to other stimulus classes. One such example is the experiments conducted by Kinnell and Dennis (2012), who found significant effects of list length on recognition performance for non-word stimuli (namely face and fractal images) while employing all of the controls for list length designs that have consistently produced null list length effects with word stimuli (Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011). Kinnell and Dennis (2012) argued that positive list length effects for novel non-word stimuli suggest that these stimuli are vulnerable to item noise at retrieval, which would fall naturally from a global matching model that has overlapping representations for these stimulus classes. The present investigation aimed to evaluate whether the strength of the studied items is an additional source of interference, which is another prediction from the first wave of global matching models. We tested this hypothesis by using fractal images, face images, and natural scene photographs in a list strength paradigm that employs the list length controls advocated by Dennis and colleagues to ensure that the finding of a positive list strength effect would not be artefactual.

In order to facilitate comparison of our results to those of the list length studies of Kinnell and Dennis (2012) which employed the same stimuli, the effect sizes (as measured by η_p^2) are plotted for each of the list strength experiments in our investigation along with the previously published list length experiments and can be seen in Figure 7. In order to estimate the standard error of the effect sizes, a bootstrap procedure was performed in which the data from each participant was sampled randomly with replacement to generate a total of 10,000 simulated datasets. η_p^2 was calculated on the difference in performance measures between each of the conditions (d' for the yes/no experiments, $p(c)$ for the 2AFC experiments) and the standard error is the standard deviation across all of the bootstrapped η_p^2 values (Efron & Tibshirani, 1985). One should note that the error estimates on the effect sizes for the list length experiments are larger in some cases because our experiments contained a much larger number of participants.

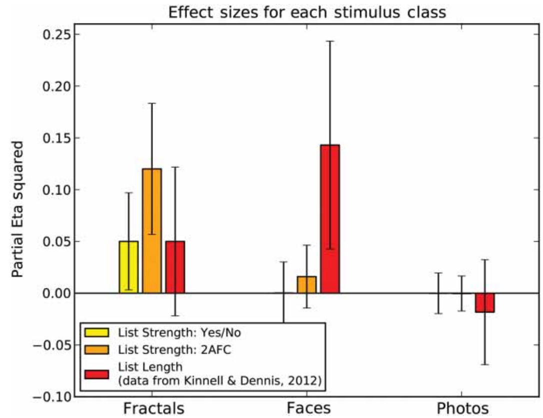

Figure 7. *Effect sizes, as measured by η_p^2, for each stimulus class in the three presented list strength experiments along with the list length experiments using the same stimulus classes conducted by Kinnell and Dennis (2012). Error bars reflect the standard error of the effect size, which was calculated from a bootstrap procedure (see the text for details).*

Results of our experiments indicated a positive list strength effect for fractal images using both yes/no and 2AFC testing, while null list strength effects were found for face images and natural scene photographs in both testing methods. Taken together with the list length experiments of Kinnell and Dennis (2012), the results for fractal images suggest that exemplars from this stimulus class can be well accommodated by a simple global matching model that utilizes overlapping representations for fractal images, making them susceptible to the effects of item noise.

Natural scene photographs exhibited null effects of both list length and list strength, which is qualitatively consistent with the results found for word stimuli and suggests that the two stimulusclasses can be well accommodated by a model that utilizes unitized representations for each of these classes. It might be reasonable to assume that natural scene photographs have unitized representations, as participants still perform quite well in recognition tests when the test stimuli are verbal descriptions of the studied photos (Paivio, 1976; Madigan, 1983). Additionally, both pictures and words rated as familiar show worse performance on recognition memory tests relative to stimuli rated as unfamiliar and show superior performance in tests of recall (Karlsen & Snodgrass, 2004).

Face images exhibited no effect of list strength, which is in contrast to the positive list length effect in the previous investigation. While our hypothesis of face images possessing overlapping representations predicts positive effects of list length and list strength on recognition memory performance, we would also like to emphasize that there is a possibility that list strength manipulations induce less interference than list length manipulations. Specifically, it is unlikely that each presentation of a repeated item was encoded with the same strength. As we previously mentioned, several computational models possess mechanisms whereby redundant stimuli are encoded with less efficacy than repeated stimuli (Farrell & Lewandowsky, 2002; Lewandowsky & Murdock, 1989; McDowd & Murdock, 1986; Murdock, 2003). This mechanism makes it such that item repetitions yield diminishing returns in encoding strength with each repetition and might predict less interference from a list strength manipulation than a list length manipulation. Thus, it is possible that face images are subject to list strength effects, especially given that a positive list strength effect was found with artificially generated face images by Norman et al. (2008).

Nonetheless, all item noise effects found in our experiments and the studies of Kinnell and Dennis (2012) were rather small in magnitude. If the only interference that these stimuli suffered came from memories from the list episode alone, it would be expected that effects of list length and list strength might be considerably larger than what we observed. One possibility is that overlapping representations do not just cause a stimulus to suffer interference from similar stimuli acquired during the list episode, but they also cause the stimulus to suffer interference from similar stimuli prior to the list episode. We refer to interference from memories that mismatch in both item and context information as *background noise*. If there is a large amount of background noise in the memory system, additional interference from item noise manipulations such as a list length or list strength manipulation would cause only a negligible increase in interference. Such an explanation follows naturally from a global matching model when memories from prior list contexts are included in the contents of memory, and was used by Murdock and Kahana (1993) to demonstrate that a null effect of list strength could be predicted in this manner. While there are likely to be many factors that are relevant to understanding the distinctions between words and non-words in recognition memory performance, such as encoding efficacy (Xu & Malmberg, 2007) and representational features that denote the stimulus class (Criss & Shiffrin, 2005), the idea that certain non-word stimulus classes suffer from greater item noise and background noise than word stimuli might be an additional candidate for understanding differences between the stimulus classes.

Another question remains as to why words might possess unitized representations while non-word stimulus classes possess overlapping representations. While we lack a formal model that describes these representational differences,[3] one possibility is that item representations begin as an overlapping representation of perceptual features

[3]Nelson and Shiffrin (2013) presented a model that describes how item representations evolve with experience, although their model does not unitize representations in the way that we hypothesize.

and through experience become bound into a relatively unitized linguistic representation. A unitized representation makes it such that a stimulus is not similar to the other contents of memory, exempting it from both item and background noise at retrieval. However, such a stimulus representation still matches its own prior representations in memory, leaving it susceptible to context noise. A similar idea was proposed by Reder, Angstadt, Cary, Erickson, and Ayers (2002) to explain the non-monotonic relationship between word frequency and recognition memory performance: very low frequency words perform worse than low frequency words because of their overlapping representations relative to the unitized representations of the low frequency words, while low frequency words perform better than high frequency words because the additional contextual associations of the high frequency words degrades performance. It is possible that the acquisition of meaning for a particular stimulus may be a component of unitization, as the training studies of Chalmers and Humphreys (1998) found that very low frequency words that were trained with their definitions exhibited a positive correlation between training frequency and performance in a subsequent recognition memory test. Words that were trained without definitions, in contrast, exhibited an inverted-U shaped relation between training and recognition memory performance.

In addition to the list strength effects found using fractal images, an additional relevant finding was that all of these experiments demonstrated more conservative responding in the mixed list relative to the pure weak list (higher values of c). This provides further support for the hypothesis of Hirshman (1995) that increases in conservative responding with strength is a regularity of the list strength paradigm, even with stimulus classes other than words. All experiments also exhibited a *strength-based mirror effect* (Stretch & Wixted, 1998): false alarm rates were lower in conditions of higher list strength, which accompanied the large increase in the hit rate for strong items relative to weak items. In some cases, the false alarm rate reductions were rather large in magnitude. In the fractal images experiment, the false alarm rates were .13 lower in the mixed condition ($\eta_p^2 = .40$) and in the faces experiment, the false alarm rates were lower by .12 ($\eta_p^2 = .39$).

Differentiation models of episodic memory

Up to this point, we have primarily restricted our discussion of the results with reference to simple global matching models and postulated that the differences between the stimulus classes can be explained in terms of differences in the degree of representational overlap. However, inferences about representations depend upon the choice of process model that is considered by the theorist (Tulving & Bower, 1974), and if the results were considered from the perspective of other process models, different conclusions about the stimulus class representations could be reached. While global matching has been a dominant view of retrieval in episodic recognition, more complex global matching models referred to as *differentiation models* give a different account of the null list strength effect in recognition memory. Differentiation models include a variant of the Search of Associative Memory model (SAM; Shiffrin et al., 1990), along with the REM (Shiffrin & Steyvers, 1997) and SLiM (McClelland & Chappell, 1998) models. Differentiation models may be able to account for some of the results we consider without choosing different representational assumptions for the stimulus classes.

In differentiation models, repetitions of an item are not stored separately, but instead accumulate into a single strong memory trace that is more responsive to its own cue and less responsive to other cues. Thus, as items are strengthened, stored item representations exhibit less of a match to non-target probe cues. Functionally, this can be described as a decrease in item noise as the strength of the items is increased (predicting a null or slightly negative list strength effect), whereas the original global matching models exhibited an increase in item noise with strength (predicting a positive list strength effect). One of the original motivations for the differentiation mechanism cited by Shiffrin et al. (1990) is that it enables global matching models to predict a positive list length effect while simultaneously predicting a null list strength effect. We refer to this key prediction of differentiation models as the *list length/list strength dissociation*.

Another consequence of the reduced item noise in conditions of higher list strength is that distracter cues will elicit less memory strength when a stronger list is studied, reducing the FAR. Thus, differentiation provides a mechanism for predicting the strength-based mirror effect without recourse to a criterion shift. Criss (2006) demonstrated that in REM, when differentiation is eliminated a list strength effect is predicted and the strength-based mirror effect is broken: FAR are predicted to be higher in conditions of higher list strength.

While these experiments were not intended to discriminate between differentiation models and a simple global matching model that employs different degrees of representational overlap for different stimulus classes, the adequacy of the differentiation account can be considered on several grounds with reference to the results presented in Figure 7. First, the predicted list strength/list length dissociation was observed for face images, in that they exhibited a significant list length effectwhile not exhibiting any effect of list strength and also exhibited a strength-based mirror effect. Thus, differentiation models provide a better account of the faces data than the model we propose, without recourse to different representational assumptions for faces. The simple global matching model we propose instead states that there should be a positive correlation between list length and list strength effects.

Nonetheless, the predicted list length/list strength dissociation is not observed for fractal images, which were subject to both list length and list strength effects, along with the natural scene photographs, which exhibited null effects of list length and list strength. While differentiation models could account for the fractals result by positing that the differentiation occurs with face images but not with fractal images, such a model would predict that the strength-based mirror effect should be broken and higher FAR should be observed in conditions of higher list strength for fractal images. The data, in contrast, showed a robust strength-based mirror effect for fractal images with significantly lower FAR in the mixed list condition. Nonetheless, these data are not strong evidence against differentiation models, and a stronger test could be devised by running an experiment manipulating list length and list strength within subjects using all of the controls in the present investigation.

CONCLUSIONS

Kinnell and Dennis (2012) found evidence for item noise in non-word stimuli when they were tested in a list length paradigm. Using the same stimuli and experimental parameters, we attempted to see whether or not positive effects of list strength could be found for non-word stimuli. While we found positive list strength effects for fractal images, the effects were small in magnitude and comparable to the effect sizes reported in the list length experiments of Kinnell and Dennis, suggesting that item noise may not be the bulk of interference in recognition memory designs as the global matching models might suggest. We have posited that background noise from the memory system may be relevant to understanding differences in performance between stimulus classes.

Original manuscript received 15 Aug 2013
Accepted revision received 11 Nov 2013
First published online 3 March 2014

REFERENCES

Chalmers, K. A., & Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 610–632.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and 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., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316–326.

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination in associative recognition and implications for representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1199–1212.

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.

Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behav- iormetrika*, *17*, 1–35.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin and Review*, *9*, 59–79.

Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., & Kahana, M. J. (2007). PyEPL: A cross-platform experiment-programming library. *Behavior Research Methods*, *39*, 950–958.

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

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 500–513.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list- strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 302–313.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix and TODAM models. *Journal of Mathematical Psychology*, *33*, 36–67.

Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 933–953.

Karlsen, P. J., & Snodgrass, J. G. (2004). The word-frequency paradox for recall/recognition occurs for pictures. *Psychological Research*, *68*, 271–276.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition*, *39*, 348–363.

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition*, *40*, 311–325.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*, 25–57.

Madigan, S. (1983). Picture memory. In J. C. Yuillie (Ed.), *Imagery, memory, and cognition: Essays in honor of Allan Paivio* (pp. 65–89). Hillsdale, NJ: Lawrence Erlbaum.

Martinez, A. M., & Benavente, R. (1998). *The AR Face Database* (Tech. Rep. No. 24). CVC.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–760.

McDowd, J. M., & Murdock, B. B. (1986). Mathematical models of memory and the problem of stimulus variation: A comparison of MINERVA 2 and TODAM. *Acta Psychologica*, *62*, 177–188.

Murdock, B. B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review*, *10*, 570–588.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 689–697.

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

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition*, *19*, 119–130.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*, 356–394.

Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin & Review*, *15*, 36–43.

Paivio, A. (1971). *Imagery and verbal processes*. London: Holt, Rinehart, & Winston.

Paivio, A. (1976). Imagery in recall and recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 103–129). New York, NY: Wiley.

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*, 163–178.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory: Receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763–785.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190–214.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 138–152.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*, 389–401.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156–163.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 267–287.

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

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207–222.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*, 1–34.

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

Strong, E. K. J. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*, 447–462.

Tulving, E., & Bower, G. H. (1974). The logic of memory representations. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 265–301). New York, NY: Academic Press.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600.

Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal and nonverbal pair strength on associative recognition. *Memory & Cognition*, *35*, 526–544.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 345–355.