Category-length and category-strength effects using images of scenes

Baumann, Oliver; Vromen, Joyce M.G.; Boddy, Adam C.; Crawshaw, Eloise; Humphreys, Michael S.

Published in:
Memory and Cognition

DOI:
10.3758/s13421-018-0833-5

Published: 01/11/2018

Document Version:
Peer reviewed version

Title: Category-Length and Category-Strength Effects Using Images of Scenes

Authors: Oliver Baumann[1,2], Joyce M.G. Vromen[1], Adam C Boddy[1], Eloise Crawshaw[1], Michael S. Humphreys[3]

1. The University of Queensland, Queensland Brain Institute, St Lucia 4072, Australia

2. Bond University, School of Psychology and Interdisciplinary Centre for the Artificial Mind (iCAM), Robina 4226, Australia

3. The University of Queensland, School of Psychology, St Lucia 4072, Australia

Correspondence should be addressed to:
Oliver Baumann
Queensland Brain Institute
The University of Queensland
St Lucia, Queensland, 4072
Australia

Tel. +61 7 3346 7275
email: o.baumann@uq.edu.au

Number of Figures/Tables: 5/0

Number of Pages: 41

Abstract

Global Matching Models have provided an important theoretical framework for recognition memory. Key predictions of this class of models are that (1) increasing the number of occurrences in a study list of some items affects the performance on other items (list-strength effect), and that (2) adding new items results in a deterioration of performance on the other items (list-length effect). Experimental confirmation of these predictions has been difficult and the results have been inconsistent. A review of the existing literature, however, suggests that robust length and strength effects do occur when sufficiently similar hard to label items are used. In an effort to investigate this further we had participants study lists containing one or more members of visual scene categories (bathrooms, beaches, etc.). Experiments 1 and 2 replicated and extended previous findings showing that the study of additional category members decreased accuracy providing confirmation of the category length effect. Experiment 3 showed that repeating some category members decreased the accuracy of non-repeated members providing evidence for a category strength effect. Experiment 4 eliminated a potential challenge to these results. Taken together, these findings provide robust support for Global Matching Models of recognition memory. The overall list lengths, the category sizes and number of repetitions used demonstrated that scene categories were well suited to test the fundamental assumptions of the Global Matching Models. These include: A) interference from memories for similar items and contexts, (B) non-destructive interference, and (C) making conjunctive information available through a matching operation.

Keywords: category length; category strength, list length; list strength, global matching; item noise; recognition memory

Category-Length and Category-Strength Effects Using Images of Scenes

Global Matching Models have provided an important theoretical framework for understanding how we recognize previously encountered events (i.e. recognition memory). Many different variations of Global Matching Models have been proposed (Anderson, Silverstein, Ritz, & Jones, 1977; Eich, 1982; Gillund & Shiffrin, 1984; Hintzman, 1984; Humphreys, Bain, & Pike, 1989; Murdock, 1982; Pike, 1984). They all have in common that during recognition, target items are compared to all items in memory (globally), rather than only to a locally-stored representation (e.g. a single node in an associative network). This leads to two testable key predictions of this class of memory models. Firstly, that memory performance decreases when other items are added to memory (the list-length effect). Secondly, the model predicts that in a mixed list where some items were strengthened (e.g., presented five times), the non-strengthened items would be less well recognized than items presented the same number of times in a pure list (no strengthened items) (the list-strength effect).

The idea that all the items in the study list contribute to the signal which is used to decide whether a test item was old or new (the matching strength) is the item noise assumption and at first this assumption seemed well supported (Clark & Gronlund, 1996). It explained the reduction in recognition accuracy when the length of the study list increased. That is, in these models as additional items were added to the study list the difference between the old and new item matching strengths remained constant. However, the variance of the matching strengths increased so that discriminability (d' or forced-choice accuracy) decreased. Increased variance, plus the assumption that items within a taxonomic category are more similar to each other than they are to items in a different category, also explained the decrease in discriminability as the number of semantically or associatively related items within a longer list increases (the category length effect) (Arndt & Hirshman, 1998; Hintzman, 1988; Shiffrin, Huber, & Marinelli, 1995).

Ratcliff, Clark, and Shiffrin (1990) also noted that the Global Matching Models predicted a list-strength effect. Although they found a list-strength effect in free recall and possibly in cued recall there was no evidence for a list-strength effect in recognition. Shiffrin, Ratcliff, and Clark (1990) noted that the null list-strength effect invalidated all of the Global Matching Models. Shiffrin and colleagues proposed, however, that the models which stored memories separately (SAM and Minerva II) could be modified to account for the null list-strength effect, by assuming that repeating items induces them to become less similar to other items in memory. Although the null list-strength effect in recognition has been replicated several times there are a small number of exceptions. Both Norman (2002) and Burrato and Lamberts (2008) have found list-strength effects in a plurality discrimination task. In this task participants can reject the opposite plurality distractors if they cue the recall of the studied opposite plurality word. Thus, the list-strength effects in plurality discrimination tasks may be due to the deliberate use of a recall strategy and therefore not truly representative of other recognition memory tasks. Furthermore, Norman (2002) had his participants make size judgments on words presented at a relatively fast rate and found a list-strength effect during recognition. The results suggest that if words, due to time limitations, are processed visually rather than semantically strength effects can be observed.

As Dennis and Humphreys (2001) noted, the weakness of the Shiffrin et al. (1990) position suggesting that the unmodified Global Matching Models needed to be rejected, was the assumption that there was a substantial positive list-length effect. Dennis and Humphreys pointed out several confounds in the studies that had purported to find list-length effects in recognition. The most important of these confounds was that the length of the retention interval

was typically correlated with lists' length. When they corrected these confounds they found very small and non-significant list-length effects. As a result, they proposed that recognition memory is not affected by interference from other items. Instead, they proposed that errors are due to noise generated by prior contexts in which the test item appeared (the context noise assumption). Subsequent research has largely supported the finding that list-length effects with words are at best quite small (Buratto and Lamberts, 2008; Dennis, Lee, & Kinnell 2008; Kinnell & Dennis, 2011; but see Cary & Reder, 2003 and Jang & Huber, 2008).

In addition, the category length effect and even the assumption that words which are members of the same taxonomic category have similar representations have been challenged. Maguire, Humphreys, Dennis, and Lee (2010) reported on two experiments using conventional methods, i.e. typical taxonomic categories, in which members were presented in a distributive fashion in the study list. They used two alternative forced choice (2AFC) tests with category lengths of one and five. In both experiments, they found no effect of length. These results have been replicated by Cho and Neely (2013) with similar procedures and materials using both a 2AFC test and a yes/no task with confidence ratings, and varying category length (i.e. 2, 8, and 14). They reported a significant effect of length when using d' but not when using the more appropriate d_a measure which controls for a difference in response criterion. These results contrasted with the results reported by Neely and Tse (2009) where the blocked presentation of category members produced a non-monotonic category length effect.

There have been, however, reports of list-length, list-strength and category-length effects in recognition tasks using non-verbal stimuli. Criss and Shiffrin (2004) found a category length

effect using sets of faces that were similar on the dimensions of race, age, hair color, and hair style. Norman Tepe, Nyhus, and Curran (2008) found a list-strength effect using faces and Kinnell and Dennis (2012) found a list-length effect using faces. Nevertheless, Osth, Dennis, and Kinnell (2014) using the same faces did not find a list-strength effect. However, the faces used by Kinnell and Dennis were certainly less similar than the faces in the categories used by Criss and Shiffrin and probably less similar than the artificially generated faces used by Norman et al. (2008) but it remains puzzling why the Kinnell and Dennis study produced a significant list-length effect with faces whereas the Osth et al. produced a non-significant list-strength effect with the same faces. Kinnell and Dennis (2012) also looked for list-length effects with fractals and pictures of scenes and Osth et al. using the same stimuli looked at list-strength effects. With fractals both a list-length and list-strength effect were found. Neither effect was found with scenes.

Although these results are patchy there is a suggestion that interference effects can be found as the similarity of the stimuli used increases. It is also possible that naming the stimuli (easy with scenes but difficult with fractals) can reduce the effects of similarity. These observations are supported by the category length effects found for pictures of objects (Konkle, Brady, Alvarez, & Oliva, 2010a) and pictures of scenes (Konkle, Brady, Alvarez, & Oliva, 2010b). In both these studies category-length (i.e. number of exemplars per category) was manipulated using a single long study list containing multiple categories, with placement of the exemplars from each category within the long list occurring in a non-blocked fashion to disguise the categorical structure. Using a 2AFC recognition task they found modest category length effects supporting the idea that these interference effects will emerge as item similarity and naming difficulty increases. Pictures seem to lend themselves more readily to generate semantically highly similar category exemplars than words do, since it is not possible to describe exemplars

of base-level taxonomic categories using just a single word. For example, Konkle et al (2010a) employed 16 different exemplars of the taxonomic category "desk", which would not be feasible using just a single word. Instead, experiments employing words have to resort to using superordinate taxonomic categories such as "furniture", which necessarily results in greater semantic and sensory-motor dissimilarities among the category members.

Investigating category length and strength effects with object and scene categories is also important for methodological reasons. Finding list-length and list-strength effects necessarily involves a comparison of performance on items from short and long lists. Although there are solutions to the problems this causes (see Dennis & Humphreys, 2001) these solutions may not work perfectly. In contrast, category length and strength can be manipulated within a list. In addition, there are a large number of object and scene categories with large numbers of instances within each category.

However, current findings with object and scene categories are not an ideal start to investigations of category length and strength effects. The most important problem is that both the Konkle et al. (2010a) and the Konkle et al. (2010b) study were concerned with demonstrating the overall human capacity for remembering visual stimuli. As a consequence they used extremely long study lists. These long lists raise problems of inattention that could reduce the overall magnitude of the category length effects[1]. In addition, even if the stimuli in different categories provide only a tiny increment to the overall noise level, the presence of

[1] Konkle et al. included a repeat detection task to assess attention during the study phase. The results of this task suggested that inattention was not a major problem, but did not rule out the possibility that inattention may have had some influence.

thousands of these low interference items could also reduce the overall magnitude of the category length effects. Furthermore, in the experiment using scene categories it was not possible to compare performance when a single category member had been studied with performance when multiple category members had been studied. This occurred for two reasons. First, the distractor used when a single category member had been studied did not come from that category as it did when multiple category members had been studied. Second, when multiple category instances had been studied, four of the studied instances were tested. As Cho and Neely (2013) noted it is possible that there are test order effects so performance on a category instance may vary depending on the number of prior instances from that category that have been tested.

Robust evidence for or against the existence of category length and strength effects is crucial to decide whether Global Matching Models or which ones should be rejected or not. Given the promising results by Konkle et al. 2010b, we wanted to investigate whether having participants study lists containing one or more members of visual scene categories is indeed a viable way to provide evidence for category length and strength effects in recognition memory. Experiments 1 and 2 aimed to replicate and extend the findings by Konkle and colleagues by testing for the existence of category length effects, and assessing its stability over time, in the face of restudy events and with different recognition paradigms. Using the same experimental approach, Experiment 3 aimed to provide the first evidence for the existence of category strength effects. Experiment 4 then aimed to eliminate a potential challenge to the first three experiments.

Experiment 1

Experiment 1 was designed to replicate the category length effect using pictures of scenes found by Konkle et al. (2010b). It differed from the Konkle et al. (2010b) study in using a much shorter study list (330 study images as compared to 2912 study images) and compared study categories of one and ten items. Only a single study item from each category was tested.

In addition, a pilot study had suggested that forgetting over a one-week period was less for studied items from long categories than for studied items from short categories. For this reason we included a one-week retention interval in the design to assess the stability of the category length effects over time. Finally, we were also interested to investigate how category length effects are affected when target items are strengthened relative to the interfering items. In speculating about why they did not find a list strength effect using the same faces with which Kinnell and Dennis (2012) had found a list length effect, Osth et al. (2014) suggested that the amount of learning due to a repetition was not as large as the initial learning of that item. This seems possible especially if there are a large number of repeated items as in the list-strength experiments (Malmi, 1977). However, it also seems possible, even though none of the existing models predicts it, that there is less interference between different strength items than between same strength items. In an attempt to understand the role of strength in interference we introduced a strength manipulation, by having participants restudy a proportion of the target items before retrieval.

Method

Participants

Seventy-one naïve participants (51 females; age: *M* = 20 years, *SD* = 3.07) gave informed consent and completed both sessions of Experiment 1. Eight participants were excluded from the analyses because their averaged recognition performance for session 1 was at ceiling (100%) and another three participants were excluded because their averaged recognition performance for session 1 did not exceed chance level (50%). Sixty participants (44 females; age: *M* = 21 years, *SD* = 3.11 years) were included in the analyses presented below. We aimed for sixty participants for all four experiments, since our pilot data indicated that this number would give us adequate power (calculated following Potvin & Schutz, 2000) for detecting main effects of interference and retention interval (i.e. >= .99), as well as the interaction between these factors (i.e. >= .90).

Stimuli

The stimuli were 390 scene images (coloured, measuring 19 cm × 25.5 cm) from different semantic (taxonomic) scene categories (e.g. *bathroom*, *supermarket*, *desert*, *beach*, etc.). The images were part of a total pool of 704 images that were gathered for the four experiments using Google Image Search. Of the 390 images, 330 were chosen as study items and the remaining 60 were used as distractors in the recognition phase. The study images were split into 30 high interference categories (ten images per taxonomic category, of which one was the target image to be tested and nine were the interfering items) and 30 low interference categories (one image per taxonomic category, being a target item in the recognition test). The image categories that were used in the immediate recognition phase, the restudy phase and the delayed recognition phase were counterbalanced over participants to rule out confounds due to potential differences in difficulty associated with the image categories. For both conditions, equal proportions of indoor and outdoor categories were used.

Procedure

The experiment consisted of two sessions that were separated by a 1-week retention interval. The first session included a study phase, a 10-minute retention interval, a recognition phase, and a second study phase. The second session consisted just of a recognition phase.

First study phase

During the initial study phase, participants viewed the 330 images (in equal proportions indoor and outdoor scenes) and were asked to memorise them as accurately as possible. Images were presented one at a time (subtending 18° × 24° visual angle) for 5 seconds each followed by a 600 ms break (see Fig. 1a). For the thirty low interference scene categories, the single image from that category was presented, which would be also a target image in the recognition phase. For the thirty high interference scene categories, each of the ten images was presented, of which one was chosen as the target image for the recognition test. Images were presented in a randomized order for each participant, (i.e. no category blocks) to disguise the categorical structure. The study phase took approximately 35-minutes to complete.

First recognition test

After a 10-minute retention interval during which participants read a book excerpt, they completed the first recognition test phase. This test phase consisted of 20 2AFC trials. On each trial, two images from the same scene category were presented horizontally next to each other – one was a previously studied target image and the other a distractor image that participants had not seen before (see Figure 1b). Participants were instructed to indicate which of the two

scene pictures they had previously studied. No feedback was provided. In half of the trials the target image was on the left side of the screen and the distractor on the right, and vice versa for the other half of the trials. Half of the trials (i.e. 10) contained images from the low interference scene categories and the other half contained images from the high interference scene categories. The presentation order of the stimulus pairs was randomized for each participant. Participants proceeded at their own pace and were told to emphasize accuracy, not speed.

Second study phase

Following the first test phase, participants were shown a selection of 20 images (in equal proportions indoor and outdoor scenes) for a second time, 10 images from the high interference and low interference categories each. These images were presented in the same manner as during the initial study phase.

Delayed recognition test

Exactly seven days after the first session, participants completed a second recognition test. This test consisted of 40 2AFC trials. Again, half of the trials (i.e. 20) contained a previously seen image from a low interference category and the other half a picture from a high interference category. Moreover, half of the trials contained a picture that had been studied twice (i.e. during the initial study phase and the restudy phase), whereas the other half of the trials contained a picture that was studied only once (during the initial study phase). Consequently, there were 10 trials with images from a high interference category that were studied once, 10 from a low interference category image that were studied once, 10 from a high interference category that were studied twice, and 10 from a low interference category that was studied twice. As mentioned above, we further counterbalanced over participants which image categories and

exemplars were used in the immediate recognition phase, the restudy phase and the delayed recognition phase.

----------Insert Figure 1 about here----------

Results

Performance in the immediate and delayed tests (10 minutes and 7 days after study) is plotted in Figure 2. Firstly, to reassess the effect of interference on recognition and to test whether forgetting over a 1-week retention is less in high interference than in low interference situations we employed a 2x2 repeated measures general linear model, analysing only non-repeat items. We found significantly lower performance for high interference than low interference study items ($F(1,59) = 12.3$, $p = .001$, *partial-*$\eta^2 = .173$) therefore replicating the effect observed by Konkle et al (2010b). We also found a significant main effect of retention interval (i.e. performance was lower after a week compared to the immediate test, $F(1,59) = 40.1$, $p < .001$, *partial-*$\eta^2 = .404$). However, we did not find a significant interaction between the effect of the retention interval and the degree of interference ($F(1,59) = 0.0$, $p = .875$, *partial-*$\eta^2 = .000$), therefore indicating, opposite to the results of the pilot study, that the rate of forgetting does not vary under high compared to low interference situations.

Secondly, to test the effect of item strength in interference we ran another 2x2 repeated measures general linear model on the performance for the delayed recognition test. We found a significant strengthening effect (i.e. better recognition for items studied twice compared to once, $F(1,59) = 97.0$, $p = < .001$, *partial-*$\eta^2 = .622$), as well as again a main effect of better

recognition for low interference than high interference situations ($F(1,59)$ = (5.7, p = .020, *partial-*η^2 = .088). In addition, we observed a significant interaction between the two factors ($F(1,59)$ = 4.2,p = 0.046, *partial-*η^2 = .066), indicating that restudying items significantly reduced the advantage in performance of low compared to high interference situations (see the rightmost four bars in Figure 2). To confirm this observation we conducted a series of three paired t-tests (using Bonferroni adjusted alpha levels of .0167 per test) comparing performance for high and low interference items for the immediate recognition test, as well as for the delayed recognition test, separately for items studied once and twice. In the immediate test, participants correctly identified significantly fewer images from the high interference category than from the low interference category (76% vs. 83%, SEMs = 2%, $t(59) = 2.61$, p = .011, Cohen's d = 0.337.). Reduced recognition accuracy for pictures from the high interference category compared to the low interference category was also observed in the delayed test for images that were only studied once (67% vs. 74%, SEMs = 2%, $t(59) = 2.80$, p = .007, Cohen's d = 0.363). For images that were studied twice, identification accuracy on the delayed recognition test was not statistically significant different in accuracy between the high and low interference items (84% vs. 86%, SEMs = 2%, $t(59) = .78$, p = .440, Cohen's d = 0.100).

----------Insert Figure 2 about here----------

Discussion

Recognition accuracy on a 2AFC test declined from the condition where only the target in a category had been studied to the condition where the target plus nine other category members had been studied. This occurred both on an immediate test and on a test delayed by one week with no indication that forgetting was less in the high interference condition. We have thus

replicated and extended the Konkle et al. (2010b) findings of a category length effect with images of scenes. However, the category length effect on the one week delayed test largely disappeared when an extra study trial was introduced. This does not appear to be a ceiling effect and may indicate that there is something special about the interference that occurs when the interfered item and the interfering items have the same strength.

Experiment 2

As mentioned above, in their review of category length effects using words, which were members of taxonomic categories and presented in a distributed fashion, Cho and Neely (2013) concluded that category length effects were not found with a 2AFC task and when a signal detection analysis allowed for unequal variances. Shiffrin et al. (1995) used a rating task and although they did not report values of d_a Cho and Neely (2013) reported that the Receiver Operating Characteristic (ROC) curves provided in an appendix suggested that there would be a significant difference in d_a. However, they also pointed out that Shiffrin et al. (1995) had not used typical taxonomic categories, but instead created categories by selecting words they judged to be semantically related to a cue word. In fact, casual inspection suggests that several of the selected words appear in the University of South Florida word association norms (Nelson, McEvoy, & Schreiber, 1998). Cho and Neely (2013) considered them as associative categories which are known to produce category length effects in a forced choice task (Maguire et al., 2010). For discussions of alternative explanations as to why associative categories and blocked presentations might produce category length effects when taxonomic categories with distributed presentations do not, we refer the reader to Cho & Neely (2013), Maguire et al. (2010) and Neely & Tse (2009). Given these mixed results and differing theories about what produces a category length effect we conducted a confidence-rating task which permits us to

compute the area under the ROC curve (this should be equal to the probability correct in a 2AFC task) as well as calculating the parameters of the unequal signal detection model. If we find a category length effect with this task it will demonstrate that the category length effects observed in Experiment 1 are not specific to a 2AFC recognition paradigm. Furthermore, it will increase confidence in the proposition that the words used in the tasks reviewed by Cho and Neely (2013) are not as similar as the scenes that we and Konkle et al. (2010) used.

Method

Participants

Sixty-three naïve participants gave informed consent and completed Experiment 2. Three participants were excluded from the analyses because their averaged recognition performance did not exceed chance level (50% accuracy). Thus 60 participants (41 females; age: M = 21 years, SD = 5.77 years) were included in the analyses presented below.

Stimuli

The stimuli were 440 scene images (coloured, measuring 19 cm × 25.5 cm) from different semantic (taxonomic) scene categories (e.g. *bathroom*, *supermarket*, *desert*, *beach*, etc.). The images were part of the same pool of 704 images that was used in Experiment 1. For each participant, 220 images were selected as study items and 40 were used as distractors in the recognition phase. Please note that the remaining 180 images were necessary to fully counterbalance the assignment of image categories to the high interference and low interference categories over participants, to avoid confounds due to potential differences in difficulty associated with the image categories. The study images were split into 20 high

interference categories (ten images per taxonomic category, of which one was the target image to be tested and nine were the interfering items) and 20 low interference categories (one image per taxonomic category, being a target item in the recognition test). For both conditions, equal proportions of indoor and outdoor categories were used.

Procedure

The experiment consisted of single session that included a study phase, a 10-minute retention interval, and a recognition phase.

Study phase

During the study phase, participants viewed 220 images (in equal proportions indoor and outdoor scenes) and were asked to memorise them as accurately as possible. Images were presented one at a time (subtending 18° × 24° visual angle) for 5 seconds each. For the twenty low interference scene categories, the single image from that category was presented, which would be also a target image in the recognition phase. For the twenty high interference scene categories, each of the ten images was presented, of which one was chosen as the target image for the recognition test. Images were presented in a randomized order for each participant, (i.e. no category blocks) to disguise the categorical structure. The study phase took approximately 25-minutes to complete.

Recognition test

After a 10-minute retention interval during which participants performed an irrelevant buffer task (i.e. a global-local visual preference task adapted from Kimchi & Palmer (1982)), participants commenced the test phase. The test phase consisted of 80 trials in which particpants indicated their confidence as to whether they had previously seen an image. Participants were presented a target or foil at each trial and indicated their degree of recognition on a 4-point scale where "1" indicated high confidence the image was studied; "2" indicated low confidence the image was studied; "3" indicated low confidence the image was not studied, and; "4" indicated high confidence the image was not studied. Participants indicated their response by selecting the corresponding number key located above the letters on a standard keyboard. No feedback was provided. For each image category, the target image as well as a distractor from the same taxonomic category were presented. The presentation order of targets and distractors was randomized and only the response to the first of those image pairs was scored to avoid any impact of the first decision about a category member on the decision about the second member. Half of the trials (40) contained images from the low interference condition and the other half contained images from the high interference condition.

Results

Two paired t-tests showed that while false alarm rates were significantly elevated in the high interference compared to the low interference condition (0.38 vs. 0.12 SEMs 0.03 and 0.01, $t(59)=11.57$, $p<0.001$, Cohen's $d=1.744$, the hit rate was not significantly elevated in the high compared to the low interference condition (0.72 vs. 0.71, SEMs 0.02, $t(59)=0.78$, $p=0.440$, Cohen's $d=0.100$).

Given that familiarity distributions of targets and foils are known to violate the assumption of equal variance (Grider & Malmber, 2008; Cho & Neely, 2013) we calculated the bias free discriminability index d_a (see MacMillan and Creelman, 2005). We calculated d_a individually

for each participant and condition (using a maximum-likelihood estimation approach assuming a Gaussian signal detection model, implemented via the RscorePlus program (Version 5.8.3). RscorePlus is based on the method of scoring algorithm for fitting a nonlinear function to data (Dorfman & Alf, 1969; Dorfman, Beavers, & Saslow, 1973). It uses singular value decomposition, combined with a variation of the Marquardt method for nonlinear least-squares regression (Marquardt, 1963; Press, Teukolsky, Vetterling, & Flannery, 2002, 2007), to find the maximum-likelihood fit of the multiple-distribution, variable-criterion signal detection model to confidence rating-scale data. All data was corrected, by adding 1/n to each response frequency, the so-called the log-linear rule (Hautus & Lee, 1998) to avoid distortion due to confidence rating frequencies of zero. This method is superior to other solutions for problematic data including deletion or substitution, as it results in less biased estimates of sensitivity (Hautus, 1995).

Performance in the recognition task using the sensitivity measure d_a is plotted in Figure 3. Participants' sensitivity in the low interference condition (d_a = 1.32, SEM = 0.08) was significantly higher than in the high interference condition (d_a = 1.03, SEM = 0.07), as indicated by a two-tailed paired t-test, $t(59) = 3.48$, $p = .001$, Cohen's $d = 0.451$.

To further compare the results of this experiment to Experiment 1, we also computed the area under the ROC curve, which is equal to the performance expected in a 2AFC task (Green & Sweets, 1974). For this we again used a maximum-likelihood estimation approach based on a Gaussian signal detection model implemented via the RscorePlus program (5.8.3). The area under the curve analysis yielded a significantly higher accuracy in the low interference condition (81.49%, SEM = 1.25%), than in the high interference condition 73.92% (SEM = 1.46%, $t(59) = 5.52$, $p < .001$, Cohen's $d = 0.719$), and accuracies were highly similar to those obtained in study 1 (i.e. 83% vs. 76% respectively).

---------Insert Figure 3 about here----------

Discussion

Recognition accuracy on the yes/no test declined from the condition where only the target in a category had been studied to the condition where the target plus nine other category members had been studied. We have thus replicated the category length effect and shown that it occurs in a confidence-rating paradigm using non-verbal material. Evidence for the existence of a robust category length effect in recognition memory is important, given the ambiguous data so far (Cho & Neely, 2013).

Experiment 3

Several studies have investigated list strength effects in recognition but have consistently found no evidence in their favour. The idea of differentiation was introduced by Shiffrin, et al. (1990) in order to explain null list-strength effects in the presence of what were assumed to be negative list-length effects. In contrast, Dennis and Humphreys (2001) argued that with words and the appropriate controls null list-strength and null-list length effects would both be found. However, even though the existence of category-length and category-strength effects are both relevant for arguing the validity of Global Matching Models, so far as we know there has been only one study (Shiffrin et al., 1995) that investigated the existence of category-strength effects. The Shiffrin et al study used word stimuli and a forced-choice paradigm but did not find any evidence of a category strength effect. In Experiment 3 we tested for the existence of category-

strength effects, by determining whether strengthening some category exemplars decreases the likelihood that non-strengthened category exemplars would be recognized.

Method

Participants

Sixty-one naïve participants (42 females; age: *M* = 20 years, *SD* = 3. 48) gave informed consent and completed Experiment 3. Five participants were excluded from the analyses because their averaged recognition performance was at ceiling (100% accuracy) and another two participants were excluded because their averaged recognition performance did not exceed chance level (50% accuracy). Thus 54 participants (37 females; age: *M* = 20 years, *SD* = 3.64 years) were included in the analyses presented below.

Stimuli

The stimuli were 144 scene images (coloured, measuring 19 cm × 25.5 cm), 6 images each from 24 different taxonomic scene categories (in equal proportions indoor and outdoor categories). The images were of the same pool of 704 images that were used in the earlier experiments. Five images of each category were study items and the sixth a distractor item to be used in the recognition phase. One of the five study images was the target item during the recognition phase.

Procedure

The experiment consisted of single session that included a study phase, a 10-minute retention interval, and a recognition phase.

Study phase

During the study phase, participants viewed 120 images, 5 images from each scene category, and were ask to memorise them as accurately as possible. Images were presented one at a time (subtending 18° × 24° visual angle) for 5 seconds each. For the low interference scene categories (12), each of the five study images was presented once. For the high interference scene categories (12), the target image was presented only once, while the other four study images were presented three times each. Images were presented in a randomized order for each participant, (i.e. no category blocks) to disguise the categorical structure. It is important to mention that the assignment of image categories to the high interference and low interference categories was counterbalanced over participants, to avoid confounds due to potential differences in difficulty associated with the image categories. The study phase took approximately 25-minutes to complete.

Recognition test

After a 10-minute retention interval during which participants read a book excerpt, participants completed the test phase. The test phase consisted of 24 2AFC trials. As in the previous experiment, on each trial, two images from the same scene category were presented horizontally next to each other – one was a previously studied target image and the other the distractor image that participants had not seen before. In half of the trials the target image was on the left side of the screen and the distractor on the right, and vice versa for the other half of the trials. Participants were instructed to indicate which of the two scene pictures they had

previously studied. No feedback was provided. Half of the trials (12) contained a studied image from the low interference condition and the other half contained a studied image from the high interference condition. Participants proceeded at their own pace and were told to emphasize accuracy, not speed.

Results

Performance in the 2AFC task (after 10-minute retention interval) is plotted in Figure 4. Participants correctly identified on average 83% of the images (SEM = 1%). Critically, significantly fewer images from the high interference condition were correctly identified than from the low interference condition (80% vs. 85%, SEMs = 2% and 1%), as indicated by a two-tailed paired t-test, $t(53) = 2.79$, $p = .007$, Cohen's d = 0.393.

----------Insert Figure 4 about here----------

Discussion

Experiment 3 showed that repeating some category members decreased the accuracy of non-repeated members providing the first evidence for a category strength effect. Evidence for both category-length and strength effects with images of scenes strongly supports a fundamental assumption of the global matching models. Namely, that the matching strengths and variances of both targets and distractors are affected by all sufficiently similar studied items.

Experiment 4

There remains a challenge to our conclusions about finding category-length and category-strength effects with images from scene categories. By testing only one item from the multi-item categories we ensured that the number of prior items tested from a category was not affecting performance on the tests of the other items from the multi-item categories. Likewise, by randomizing the order of the study lists we insure that a target from a singleton category is, on average, studied in the same list position as the single target from a multi-item category. However, we did not control for the number of other items from the target category that had been studied prior to studying the target.[2] This never occurs with singleton categories but in general there will be several such items studied before studying the target from a multi-item category. Maguire et al. (2010) using taxonomic categories had compared performance on targets from singleton categories with performance on targets which were the first or last item studied in a multi-item category, controlling for position within the study list. They reported no systematic effects of this manipulation, which is the reason we did not attempt to control for this variable. However, failing to find an effect using category instances that do not produce a category-length effect may not apply when the category instances produce a category-length effect. For this reason, we used the Maguire et al. control procedures in Experiment 4 where we took another look at category length.

Method

Participants

[2] Konkle et al., 2010b) did not provide enough information about the four items tested from each category in their exemplar test conditions to determine whether this criticism applies to their study.

Sixty naïve participants (53 females; age: M = 21 years, SD = 6.27) gave informed consent and completed Experiment 4. No participants were excluded since the recognition performance was lower than 100% and higher than 50% for all participants.

Stimuli

The stimuli were 704 scene images (coloured, measuring 19 cm × 25.5 cm). The images were of the same pool that were used in the earlier experiments. Eleven images each were selected from 64 taxonomic scene categories (half indoor, half outdoor). For each participant 352 of those images were chosen as study items and 64 were used as distractors in the recognition phase. The study images were split into 32 high interference categories (ten images per taxonomic category, of which one was the target image to be tested and nine were the interfering items) and 32 low interference categories (1 image per taxonomic category, being also the target item during recognition).

Procedure

The experiment consisted of a single session that consisted of a study phase, a 5-minute retention interval, and a test phase.

Study phase

During the study phase, participants viewed images from the 64 scene categories. Images were again presented one at a time (subtending 18° × 24° visual angle) for 5 seconds each. For the low interference scene categories (32), a single image from that category was presented once. For the high interference scene categories (32), each of the ten images was presented. The study

phase was divided into three subphases. In the first, we presented the target images (our so-called early target condition) from 32 taxonomic scene categories (16 images each for the high and low interference conditions). In the second subphase, we presented the interfering items for the high interference conditions (9 images each for 32 taxonomic categories). In the third subphase, we presented a second set of target images (our so-called late target condition) from 32 taxonomic scene categories (16 images each for the high and low interference conditions). Therefore, half of the images presented in subphase 2 belonged to the same categories as the high interference early targets, while the other half belonged to the categories of the high interference late targets. The presentation order of the images was randomised separately for each of the three study subphases and each participant (i.e. no category blocks) to disguise the categorical structure. Further, the assignment of taxonomic categories to high versus low interference conditions and early versus late target conditions was counterbalanced over participates to rule out any confounds due to potential differences in difficulty associated with the image categories. The study phase took approximately 40-minutes to complete.

Recognition test

After the 5-minute retention interval during which participants read a book excerpt, they completed the test phase. This test phase consisted of 64 2AFC trials. As in the previous two experiments, on each trial, two images from the same scene category were presented horizontally next to each other – one was a previously studied target image and the other the distractor image that participants had not seen before. In half of the trials the target image was on the left side of the screen and the distractor on the right, and vice versa for the other half of the trials. Participants were instructed to indicate which of the two scene pictures they had previously studied. No feedback was provided. Half of the trials (32) contained a studied image

from the low interference condition and the other half contained a studied image from the high interference condition. Within the low and high interference conditions, half of the tested images early targets (i.e. amongst the first 32 image presented) and the other half were late targets (i.e. amongst the last 32 images presented). The presentation order of the test pairs was randomized for each participant. Participants proceeded at their own pace and were told to emphasize accuracy, not speed.

Results

Recognition performance is plotted in Figure 5. Firstly, to test effects of high vs. low interference situations independent of the timing when the target items were presented during the study phase we conducted a 2x2 repeated measures ANOVA. We observed significant effects of interference (high vs. low), $F(1, 59) = 29.16$, $p < .001$, *partial-*$\eta^2 = .33$, and time (early vs. late), $F(1, 59) = 8.17$, $p = .006$, *partial-*$\eta^2 = .12$. Importantly, their interaction was non-significant, $F(1, 59) = 3.06$, $p = .085$, *partial-*$\eta^2 = .05$. Two follow-up paired t-tests (using Bonferroni adjusted alpha levels of .025 per test) showed that significantly fewer images from the high interference category were correctly identified than from the low interference category, for early target presentations (81% vs. 90%, SEMs = 2% an 1%, $t(59) = 5.6$, $p < .001$, Cohen's d = 0.734), as well as late target presentations (%78 vs.83 %, SEMs = 2%, $t(59) = 2.9$, $p = .005$, Cohen's d = 0.376). As shown by the ANOVA, on average, fewer images that were presented in the third subphase of the study session as compared to the first subphase were correctly identified (81% vs. 86%, SEMs = 2% and 1% respectively.

----------Insert Figure 5 about here----------

Discussion

Performance declined from the first to the third phase of the study list. This supports Underwood's (1978) conjecture that participants get fatigued or inattentive during the course of a long study list. Although there was a slight decline in the size of the category-length effect for the last item compared to the first item, this interaction effect was not statistically significant. Importantly, recognition accuracy was significantly lower for first and last items from multi-item sets compared to singleton targets. This supports the findings from Maguire et al. (2010) and gives us high confidence that within category study order effects were not contributing to our category-length and category-strength effects.

General Discussion

The aim of this study was to investigate the use of scene taxonomic categories as a useful tool in assessing the Global Matching memory models, by determining whether the use of scenes would produce category-length and category-strength effects. In Experiments 1, 2, and 4, we have provided robust evidence for the existence of category-length effects. We also found a category-strength effect in Experiment 3. Shiffrin et al. (1990) concluded that the Global Matching Models, which assumed composite memories, could be rejected by findings of null list-strength effects. This conclusion depended on the assumption that there were robust category-length effects and was weakened by subsequent findings that showed that, with appropriate controls, category-length effects with words were small and generally non-significant. It is further weakened by our finding of a positive category-length and category-strength effect using taxonomically organised scene stimuli. The Shiffrin et al. solution to the problem was to assume that repeated items became differentiated from other items in memory.

That is, each repetition reduced the similarity between the stored image (memory) for that item and other stored images. Originally, this was a post hoc process added to the SAM model. However, Shiffrin and Steyvers (1997; also see McCelland & Chappell, 1994) incorporated the differentiation hypothesis into their assumptions about how memories are represented and stored. The current data show that this assumption is not compelling, but on the basis of current findings we cannot reject differentiation and the models that incorporate it.

In the following, we show how the global matching framework helps in evaluating our results and some key ideas about memory and forgetting. These include: 1) interference from memories for the other list items, from the other contexts in which the target item occurred, and from background memories (Osth & Dennis, 2015), (2) non-destructive interference, and (3) making conjunctive information available through a matching operation.

The first key idea is that other items and contexts constitute a source of noise in accordance with their similarity to the target item and the similarity to the context in which the target item was studied. Our experiments directly address the question as to whether the other items in the study list constitute a source of noise and indirectly address the issue of similarity. We have replicated the Konkle et al. (2010b) finding of a category-length effect using scene categories in Experiments 1, 2, and 4. Experiment 4 also eliminated a potential problem with the Konkle et al. (2010b) experiment and with our first three experiments. In addition, in Experiment 3 we found a category-strength effect. As far as we know this is the first time anybody has reported a strength effect using categories. Similarity is indirectly implicated by these results because Konkle et al. (2010b) found such good performance after subjects had studied extremely long lists (i.e. 2912 images). This suggests that scenes that are not in the same scene category do not contribute much in the way of item noise. In keeping with the global matching framework, this

suggests that scenes within a scene category are more similar to each other than are scenes in general. In addition, our review of list-length and list-strength effects suggests that the list items must be similar if these effects are to be found.

The second key idea is that the interference is non-destructive. Experiment 2 has shown that with a confidence-rating task an increase in the length of a category leads to a significant increase in the false alarm rate and no reduction in the hit rate. This suggests that a larger number of exemplars leads to an increase of overall familiarity for all category members, which in turn causes lower discriminability. The constant hit rate, however, suggests that memories have not been overwritten (also see Dyne, Humphreys, Bain, & Pike, 1990). Our Experiment 4 is also relevant to this point. In this experiment, we showed that the interfering effect of increasing the number of studied items in a category did not change as a function of whether the additional items were studied before or after the target was studied. If studying the additional category members after the target had been studied produced more interference than studying them before the target had been studied, it would have been problematic for the assumption of non-destructive interference. That is, this would have indicated a destructive form of interference, which unlike other forms of destructive interference (Wixted, 2005), would have depended on similarity.

As we have noted there are problems with list-length and list-strength manipulations because they necessarily involve comparisons between short and long lists. These problems are eliminated with category-length and category-strength manipulations because these manipulations can occur within a list. Our results with scene categories were obtained with moderate list lengths. We also used only a moderate number of interfering items (9) in our

category-length experiments and a moderate number of repetitions (2) in our category-strength experiment. Thus scene categories seem well suited to further explorations of a variety issues about the sources of noise and the causes of forgetting. In addition, they may be useful in exploring our third key idea. The Global Matching Models propose that conjunctive information ("Which list did a word occur in?", "How was the word processed?", "Did two words occur together?", etc.) can be conveyed by forming a conjunctive representation of two or more cues and matching (a computation of similarity) that conjunctive representation against the conjunctive representations in memory. The alternative is that a recall-like process commonly referred to as recollection is the primary means of conveying conjunctive information (Humphreys, 1978; Jacoby, 1993; Mandler, 1980). Humphreys and Chalmers (2016) have already reviewed evidence that conjunctive information can be conveyed by a continuous source of information and it seems likely that the use of scene categories would produce even clearer results. This is compatible with the Global Matching models, but it has been proposed that recollection might also provide a continuous source of information (Wixted & Mickes, 2010). It would also be interesting to use scene categories in associative recognition and in the Remember/Know paradigm. Presumably, participants can discriminate scenes that were studied together from scenes that were studied separately. However, it is unlikely that participants could produce a detailed image of one scene given another as a cue. Outcomes from such studies would have important implications for assumptions regarding the underlying recollection process (cf. Mather, Henkel, and Johnson; 1997).

The Global Matching framework has been extremely useful in addressing these key ideas. It has predicted new phenomena that had not been anticipated. The best example of this is the effect of strengthening some of the items in a list or category. This effect now seems well established though probably only for similar items and/or for items that are difficult to label. It

can also tell us where not to look. It is possible that the relatively small category-strength effect that we found would not have been observed if we had tried to use list lengths that approached the one used by Konkle et al. (2010b). Also, see Osth and Dennis (2015) for more examples where the effects of one of the contributing variables can obscure the effects of other variables. Hopefully the Global Matching framework is also precise enough, so it will be possible to tell when it needs to be seriously amended or perhaps discarded altogether.

References

Anderson, J. A., Silverstein, J.W., Ritz, S. A. & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 88,* 413-451.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory & Language, 39,* 371-391.

Buratto, L. G., & Lamberts, K. (2008). List strength effect without list-length effect in recognition memory. *Quarterly Journal of Experimental Psychology, 61*, 218-226.

Cary, M., & Reder, L. M. (2003) A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231–248.

Chappell, M., & Humphreys, M. (1994). An auto-associative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review, 101,* 103–128.

Cho, K. W., & Neely, J.H. (2013). Null category-length and target-lure relatedness effects in episodic recognition: a constraint on item-noise interference models. *Quarterly Journal of Experimental Psychology, 66,* 1331-1355. http.//dx.doi: 10.1080/17470218.2012.739185.

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.

Cohn M., & Moscovitch M. (2007). Dissociating measures of associative memory: Evidence and theoretical implications. *Journal of Memory and Language, 57,* 437–454.

Criss, A. H. & Shiffrin, R. M**.** Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition, 32,* 1284 - 1297.

Dennis, S., & Humphreys, M. S. (2001). The role of context in episodic recognition: The bind cue decide model of episodic memory. *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. http://dx.doi.org/10.1016/j.jml.2008.06.007

Dyne, A. M., Humphreys, M. S., Bain, J. D., & Pike, R. (1990). Associative interference effects in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16,* 813-824.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627-661.

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

Green, D. M. & Sweets, J. A. (1974). *Signal detection theory and psychophysics.* Huntington, NY: RF Krieger.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, 27(1),* 46-51.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Humphreys, M. S. (1978). Item and relational information: a case for context independent retrieval. *Journal of Verbal Learning and Verbal Behavior, 17,* 175-188.

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,* 208-233.

Humphreys, M. S., Burt, J. S., & Lawrence, S. (2001). Expecting *Dirt* but Saying *Dart*: The Creation of a Blend Memory. Psychonomic Bulletin & Review, 8, 820-826.

Humphreys, M. S., & Chalmers, K. A. (2016) *Thinking about Human Memory*. Cambridge, United Kingdom: Cambridge University Press.

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.

Humphreys, M. S., Tehan, G., O'Shea, A. & Bolland, S. W. (2000). Target similarity effects: Support for the parallel distributed processing assumptions. Memory & Cognition, 28, 798-811.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory* & *Language,* 30, 513-541.

Jang, Y. & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning Memory and Cognition, 34,* 112-127.

Kimchi, R., & Palmer, S. E. (1982). Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance, 8(4),* 521-535.

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

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

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139,* 558–578.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychological Science, 21,* 1551–1556.

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

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. J*ournal of Experimental Psychology: Human Learning and Memory*, *4,* 19-31.

Maguire, A. M., Humphreys, M. S., Dennis, S., & Lee, M. D. (2010). Global similarity accounts of embedded-category designs: Tests of the Global Matching Models. *Journal of Memory and Language, 63,* 131–148.

Malmi, R. A. (1977). Context effects in recognition memory: The frequency attribute. *Memory and Cognition, 5*, 123-130.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

Mather, M. Henkel, L. A., & Johnson, M. K. (1997). Evaluation characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25*, 826-837.

Metcalfe, J. (1990). A composite holographic associative recall model (CHARM) and blended memories eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145-160.

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

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

Neely, J. H., & Tse, C. (2009). Category length produces an inverted-U discriminability function in episodic recognition memory. Quarterly Journal of Experimental Psychology, 62, 1141–1172.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*,28, 1083–1094.

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.

Osth, A. F., Dennis, S., & Kinnel, A. (2014). Stimulus type and the list strength paradigm. *Quarterly Journal of Experimental Psychology, 67,* 1826-1841.

Ost, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*, 260-311.

Potvin, P. J., & Schutz, R. W. (2000). Statistical power for the two-factor repeated measures ANOVA. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc.*, *32*(2), 347–356.

Raaijmakers, J. G. W. & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

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

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 & 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.

Tehan, G. & Humphreys, M. S. (1998). Creating proactive interference in immediate recall: Building a dog from a dart, a mop, and a fig. Memory & Cognition, 26, 477-489.

Tehan, G., Humphreys, M. S., Tolan, G. A., & Pitcher, C. (2004). The Role of Context in Producing Item Interactions and False Memories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 107-119.

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

Verde, M. F. (2004). Associative interference in recognition memory: A dual-process account. *Memory & Cognition, 32,* 1273–1283.

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science, 14,* 6-9.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054.

Figure captions

Figure 1. Encoding and Recognition procedures. a) During encoding, participants viewed 330 images for 5 seconds each, with a 600ms inter-stimulus interval. b) During recognition, participants viewed 60 target-foil pairs (20 in the first recognition test and 40 in the delayed recognition test), and attempted to identify the image they had seen previously.

Figure 2. Percentage correct recognition (±1 SE) for the high and low interference conditions on the immediate test as well as the delayed test seven days after studying the images (* denotes a statistically significant difference between the high and low interference conditions at $p < .05$ as assessed by a two-tailed, paired t-test). Delayed test non-repeat refers to those items that were only shown once during the initial study phase, whereas delayed test repeat refers to those items that were shown a second time at the end of session one.

Figure 3. Results from Experiment 2. a) Discriminability (d_a, (±1 SE) for the high and low interference conditions. b) Area under the ROC curve for the high and low interference conditions (equivalent to the performance expected in a 2AFC task). * denotes a statistically significant difference between the high and low conditions at $p < .05$ as assessed by a two-tailed, paired t-test.

Figure 4. Percentage correct recognition (±1 SE) after 10-minute retention interval for the high and low interference conditions in Experiment 3 (* denotes a statistically significant difference between the high and low conditions at $p < .05$ as assessed by a two-tailed, paired t-test.

Figure 5. Percentage correct recognition (±1 SE) for the high and low interference conditions for images that were presented early vs late in the study phase (* denotes a statistically significant difference between the high and low interference conditions at $p < .05$ and ** at $p < .001$ as assessed by a two-tailed paired t-test.

Figure1

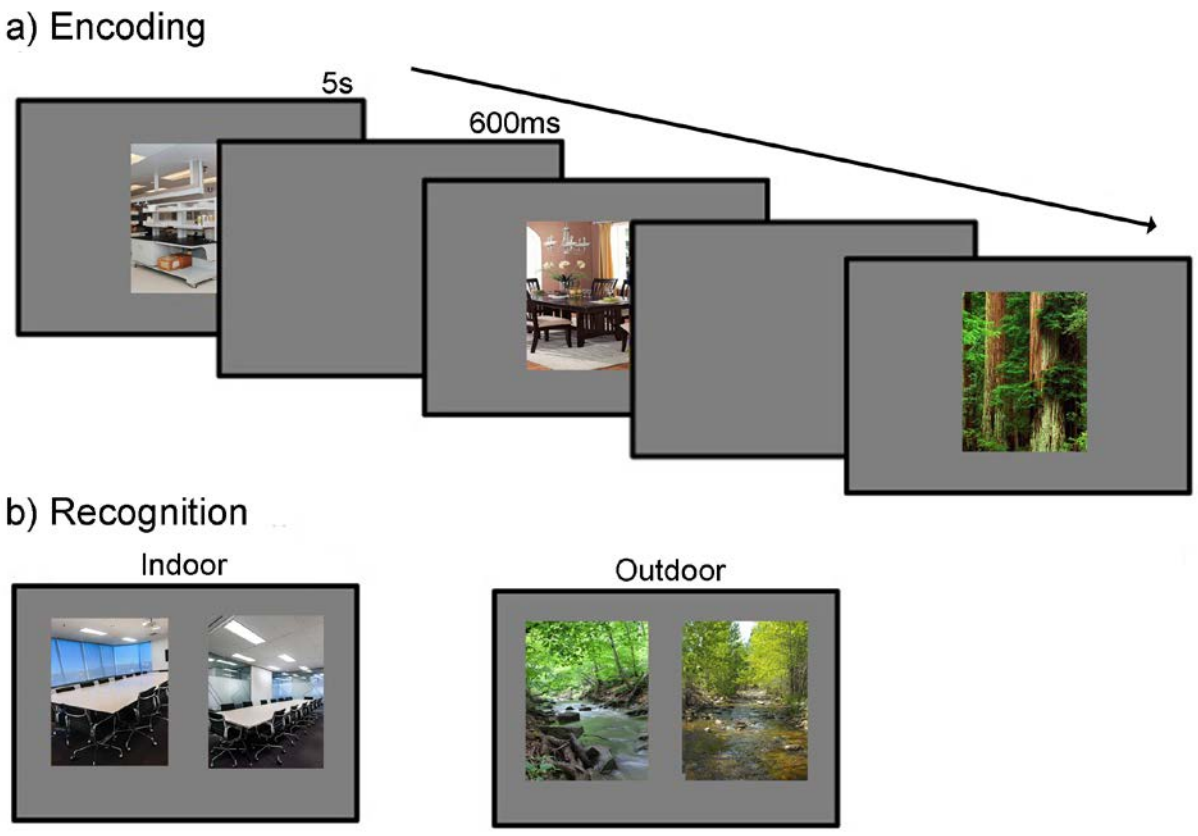

Figure 2

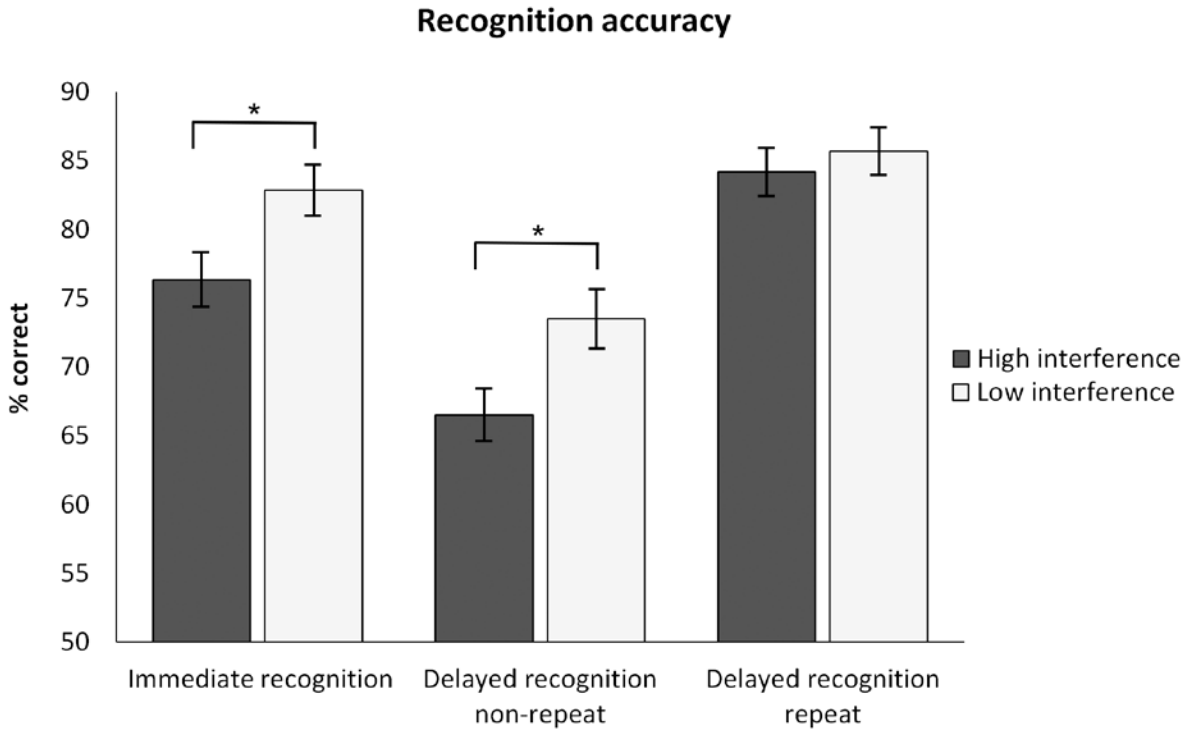

Figure 3

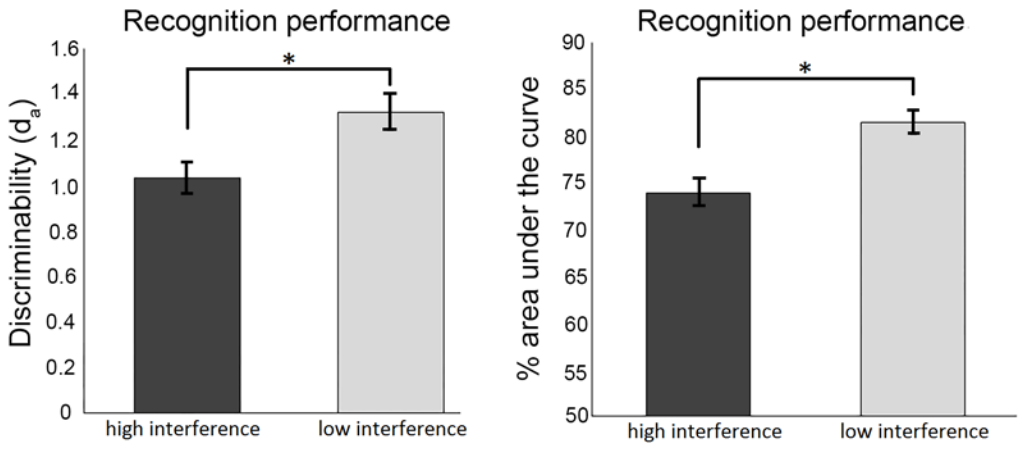

Figure 4

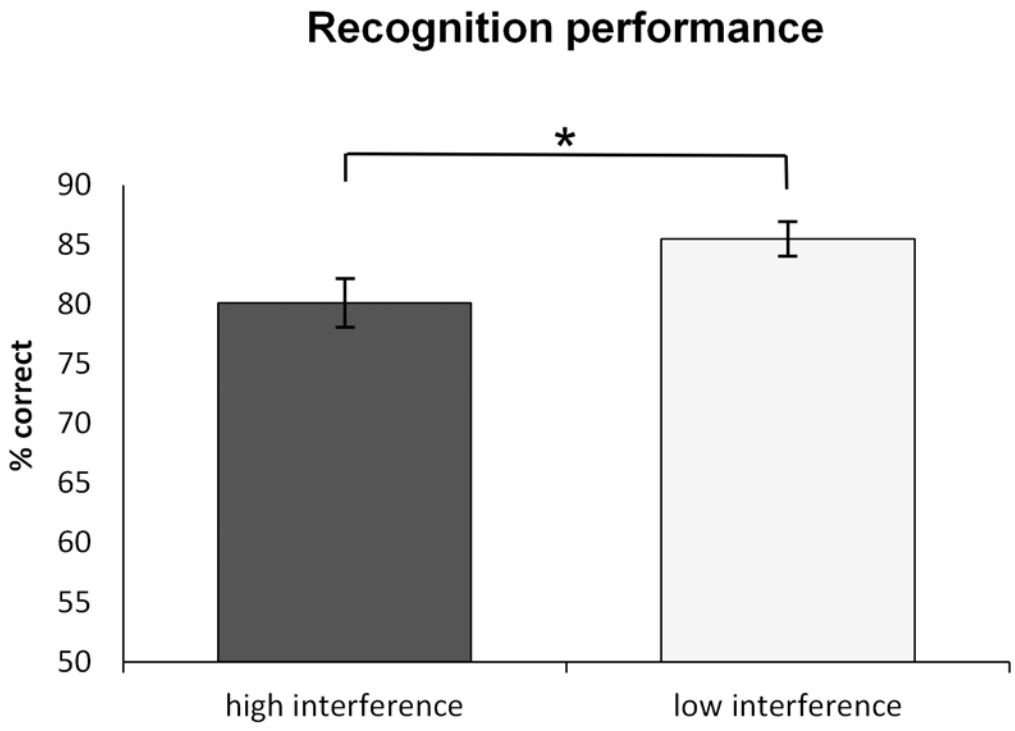

Figure 5