# Similarity and Strategic Effects in Recognition Memory

Gregory E. Cox, George Kachergis, Richard M. Shiffrin
(grcox, gkacherg, shiffrin)@indiana.edu
Department of Psychological and Brain Sciences, Cognitive Science Program, Indiana University
1101 E. Tenth St., Bloomington, IN 47405 USA

## Abstract

We introduce a class of artificial stimuli that lack preexperimental associations or encoding strategies. In a set of recognition memory experiments using these stimuli, we manipulate the similarity between studied items and between targets and foils, thus investigating the effects of pure perceptual similarity. We also assign values to studied items in order to induce encoding strategies that might emphasize encoding distinctive or overlapping features. Applying a stochastic signal detection model to these data, we find that blocked presentation and increased category size lead to poorer encoding of individual items, indicating that participants fail to encode distinctive features when list homogeneity is increased. Further, items assigned a negative value are encoded more poorly, a sign that participants may attempt to find overlapping features among negative items.

Keywords: recognition memory; categorization; similarity.

## Introduction

The manipulation of similarity between items in memory has served as a rich source of evidence for models of memory. Perhaps the most famous of these is the DRM paradigm, in which studying a list of semantically-related items leads to greater false recall and recognition of other semantically-related items (Roediger & McDermott, 1995). Within a single list, increasing the number of exemplars from a given semantic or orthographic category leads to higher false alarm rates to category members (Shiffrin, Huber, & Marinelli, 1995). And despite overall high performance, the semantic similarity among visual objects (e.g., cars, backpacks) leads to greater interference (Konkle, Brady, Alvarez, & Oliva, 2010).

The kinds of stimuli used in these experiments tend to be things with which participants have a great deal of experience, e.g., words, colors, or common objects. Because of this experience, participants come to the experiment with potentially idiosyncratic encodings or strategies (this would seem to be particularly true of verbal stimuli). This is exemplified by the classic "own-race" bias, in which face recognition is superior for member's of one's own race due to greater exposure and a corresponding ability to attend to relevant features of the face (Meissner & Brigham, 2001). As a result of these kinds of idiosyncrasies, it is sometimes difficult to make inferences about memory processes in general, since one might need to appeal to processes that are specific to the stimuli or participants employed.

In this paper, we introduce a novel class of stimuli for recognition and categorization experiments that avoids some of these problems. Because the stimuli are entirely novel and difficult to verbalize, we eliminate most effects of prior experience. They are also randomly generated for each list and each participant, minimizing interference between lists and marginalizing the effects of potentially idiosyncratic stimuli. They also allow for fine-grained parametric manipulation of inter-item similarity. Because these stimuli do not, a priori, suggest any particular encoding strategies, we can also investigate the effect of manipulating item valence on encoding, thereby implicitly making some items more "important" than others (Kachergis, Recchia, & Shiffrin, 2011). We present results from a set of recognition memory experiments in which the similarity and valence among studied items and between targets and foils is manipulated. The effects of these manipulations are interpreted within the context of a stochastic signal detection model of memory.

## Experiment 1

Episodic memory experiments typically present items with no indication of their relative importance, making it unclear why participants might devote more effort toward encoding some items rather than others. In the current experiment, we use the valence of an item—whether it is worth positive or negative points—to indicate relative importance. Some conditions contained eight positive-valued objects, while others contained four positive-valued (+10 points) and four negative-valued objects (-10 points).

We also manipulated the perceptual similarity of the eight objects in the study list: they were either all similar (i.e., 1 category), all dissimilar (8 categories), or comprised of two categories of four similar objects. Perceptual categories might or might not align with valence categories. In either case, the question is whether a homogeneous list is encoded better or worse, and whether manipulating valence leads to increased discriminability between study items. To better pull apart issues of similarity, we use different foil types during a 2-alternative forced choice (2AFC) recognition test. Foils can be from the same perceptual category as the target, from a different studied category (if there were two studied categories), or from a novel category.

### Subjects

133 undergraduates at Indiana University participated to receive course credit.

### Stimuli

Each stimulus was a light gray blob (50 pixels in diameter). Its boundary, denoted $f(\theta)$, was generated by Fourier synthesis in polar coordinates according to $f(\theta) = \sum_{i=1}^{12} i^{\frac{1}{2}} \exp\{\cos[i(\theta + \phi_i)]\}$, where $w_i$ and $\phi_i$ are the weight and phase, respectively, of the component with frequency $i$ (Shepard & Cermak, 1973). Eight stimuli were created for

(a) Category 1 (b) Category 2

Figure 1: Example blob stimuli from two categories.

Figure 2: Screenshot of a single trial, in which the participant chose an unstudied (i.e., neutral) microbe. Feedback appeared only after the decision was made.

each blob category (although less than 8 may end up in the experiment) by first randomly selecting a set of initial phases $f_i^0$, $i = 1 \ldots 12$ for each component. Then, to create 8 exemplars, the relative phase of two components, $f_3$ and $f_5$, were set to 8 equally-spaced values in the range $[0, 2\pi)$, e.g., $f_3 \in \{f_3^0, f_3^0 + \frac{\pi}{4}, f_3^0 + \frac{\pi}{2}, \ldots\}$ where $f_3^0$ is the randomly chosen initial phase, and similarly for $f_5$. Pilot studies using multidimensional scaling—not reported here due to space constraints—established that, even given the random nature of these stimuli, individual exemplars were discriminable and more similar within categories than between. Example stimuli are shown in Figure 1.

New items were generated for each participant for each of 20 blocks such that participants saw no stimulus more than once. Each study block contained eight objects, each paired with a value, either +10 or -10. Participants studied each object-value pair for four seconds, in randomized order.

## Procedure

Participants were instructed that they would be playing a game in which their goal would be to maximize their points by studying and remembering “alien microbes”, some of which are good (positive points), and some of which are bad (negative points). After studying, two microbes would fall from the top of the screen, one of which had been on the preceding study list, and they would have to choose the more valuable microbe (novel microbes were always worth zero points). At the start of each test trial, the two choice items would appear horizontally separated by 200 pixels and vertically separated from the participant's agent (which is initially equidistant between the two options) by 210 pixels, moving downward at a constant rate of 1 pixel per frame (at a refresh rate of 60 Hz) on 15” CRT monitors with a resolution of 800x600 pixels. Participants made their choice by using the arrow keys to move a small arrow-shaped agent under the microbe they wanted to choose (see Figure 2). The trial ended when the chosen microbe fell into the participant's agent or

Figure 3: Data and predictions for Experiment 1.

the participant pressed the space bar to immediately choose whichever microbe they were currently under. If the participant failed to select one of the two objects, they lost 30 points and were told to try to select one of the objects on every trial. Participants running score, tallied across all conditions, is shown throughout testing in the upper left corner of the screen.

Subjects participated in each of the ten unique study conditions twice, for a total of 20 blocks, each with eight trials. Condition order was counterbalanced across subjects.

## Results

22 participants were excluded from analysis because their overall performance was not significantly greater than chance (.531 for 160 trials). Accuracy results for the remaining 111 participants are shown in Figure 3. An accurate response in one in which the participant selects the item with greatest valence: selecting the old item if it is positive or the foil if the studied item is negative. An analysis of variance on the number of perceptual categories (1, 2, or 8), the valence composition of the study list (mixed or univalence), and the foil type (similar, other, or novel) shows significant main effects of the number of categories ($F(2,110) = 10.94$, $p < .001$), valence composition ($F(1,110) = 20.73$, $p < .001$), and foil type ($F(2,61) = 32.88$, $p < .001$). Significant interactions were: number of categories by valence ($F(2,220) = 27.80$, $p < .001$), study distribution by valence composition ($F(2,220) = 15.58$, $p < .001$), and number of categories by valence by foil type ($F(4,440) = 7.65$, $p < .01$). All other interactions had $F$-values less than one.

Accuracy in conditions with one perceptual category (i.e., all similar) was worse than accuracy in conditions with eight or two categories ($M_1 = .62$, $M_8 = .66$, $M_2 = .65$). Accuracy in conditions with only positive items was superior to accuracy in conditions with both positive and negative items ($M_{pos} = .66$; $M_{both} = .62$), but there was no significant difference in overall accuracy between positive and negative items in the mixed-valence condition ($M_{pos} = .60$, $M_{neg} = .62$; $t(110) = .81$, $p = .42$). Overall accuracy was lower when foils were similar to targets than when they were unique; however within the univalence condition, accuracy was higher for similar foils than for foils from a different category, a perplexing “similar-foil” effect originally found by

Tulving (1981).

## Experiment 2

In Exp. 2 we examined recognition memory for lists composed of either two perceptual categories, or all unique stimuli. Unlike Exp. 1, all of the stimuli were given positive values in this experiment. For lists with two categories, we examined the effect of interleaving vs. blocking the two categories during study. Prior work has shown that inductive categories are best learned from interleaved training (Carvalho & Goldstone, 2012). However, we were interested to see if more interference would come from blocking—which separates the categories in time and may lead to more prototype-like encoding—or from seeing the categories mixed, which might make it easier for participants to learn distinctive features of the items.

For the two-category lists, category was varied: equal sized (4 and 4) or unequally sized (6 and 2). More exemplars gives more opportunities to form a category representation, but with the potential cost of greater confusability. On the other hand, a small category may be better remembered due to its distinctiveness.

### Subjects

86 undergraduates at Indiana University participated to receive course credit.

### Stimuli and Procedure

The same stimuli and procedure were used as in Exp. 1.

### Design

Each study list contained 8 blobs, and participants performed 18 study-test blocks. Two blocks were comprised of unique study items (i.e., 8 categories of size 1), which were tested against either unique foils or foils that were similar to the target. There were four blocks with two studied categories (4 exemplars each). In two of these blocks, the categories were interleaved, and in the other two the categories were blocked. Finally, there were 12 blocks with two unequally-sized categories. In the two-category blocks, foils could be from the same category as the target, the other studied category, or novel.

### Results

Twelve participants were removed because their overall accuracy was not signi cantly above chance. Data from the remaining 74 participants were analyzed in terms of their probability of choosing the correct (in this case, old) item (see Figure 4). An ANOVA on category size (1, 2, 4, or 6 exemplars), list type (blocked, interleaved, or other) and foil type (similar, dissimilar, or novel) shows a signi cant main effect of foil type ($F(2,73) = 37.63$, $p < .001$)–all other F-values were less than 1. Accuracy was lower when foils were similar to targets than when they were unique, or drawn from the other category ($M_{similar} = .62$; $M_{unique} = .76$; $M_{other} = .66$). There was a signi cant interaction of category size and foil type ($F(4,146) = 3.43$, $p < .01$). The larger the category, the worse people got at discriminating similar foils from exemplars of that category, but the better they became at discriminating category members from unique foils.

Figure 4: Data and predictions for Experiment 2.

## A Model

To better understand the effects of category size, valence, and blocking/interleaving, we introduce a stochastic signal detection model. This model aims not to be a detailed process model; rather, it is hoped that the parameter estimates obtained from this model will provide a deeper understanding of the memory and decision processes that generated our data. Although this model is similar to the Generalized Context Model (GCM; Nosofsky, 1986), we do not have pairwise similarity ratings for each stimulus and subject. Therefore, we directly estimate item similarities in the model, rather than the parameters of GCM's exponential similarity rule. Further, unlike our model, GCM does not assume noise in the sensory/memory representations of item; however, stochastic noise has been shown to be critical for explaining the Tulving similar-foil effect (Hintzman, 1988; Clark, 1997). In making the assumption of stochastic noise, our model is quite similar to the NEMO model (Kahana & Sekuler, 2002).

We assume that each of the two choice items is compared to the memory traces of all eight items from the study list. Each comparison produces a match value that is proportional to both the similarity between the choice item and the memory item as well as the encoding strength of the memory item. Match values may also be weighted by the retrieved valence for each item, which may or may not have been stored correctly. The participant then selects the item with the higher summed match.

### The Match Distribution

We assume that the match value between a choice item and a memory trace is normally distributed with a mean value that depends on both the similarity between the choice item and the trace and the encoding strength of the trace. The variance of any match is assumed to be a constant, $s^2$; thus, any variation in the mean match value can be thought of as varying the signal-to-noise ratio. If there are two choice items and $N$ study items, there are then $N2$ match values which are jointly normally distributed. This joint distribution is characterized by the vector of $2N$ mean match values and the $2N \times 2N$ matrix of their covariances. Then, the distribution of the difference in summed match between the two choice items can be expressed as a linear function of the joint match distribution.

We assume the mean match value of an item to itself is 1, the mean match value between two independently generated blobs is zero, and the mean match between two blobs from the same category is $w$, $0 < w < 1$. In addition, the match values between items of the same category are positively correlated (with value $r$, $0 < r < 1$). This correlation arises from shared category features: if a choice item shares a feature with one item from category A, it is likely to share that feature with other category A items since items within a category will tend to share features. Conversely, if a choice item possesses a feature that is absent from a category A item, that feature will also tend to be absent from other category A items.

For example, say the study list consists of $N = 4$ items, with 2 items from one category and 2 items from another. If on a given trial, the foil is completely novel, the mean match vector would be $\mu = [1, w, 0, 0, 0, 0, 0, 0]^T$ and the covariances between the match values would be

$$\Sigma = \begin{bmatrix} \sigma^2 & r\sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ r\sigma^2 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & r\sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & r\sigma^2 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & r\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & r\sigma^2 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & r\sigma^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & r\sigma^2 & \sigma^2 \end{bmatrix}.$$

where the first 4 match values are matches to the target and the second 4 are matches to the foil.

The probability of selecting an old item is the probability that the difference in the summed match between the old and the new item exceeds zero. The distribution of this difference can be obtained by applying the linear operator $k = [1, 1, 1, 1, -1, -1, -1, -1]^T$ to the multivariate match distribution. This operator simply sums the target match values and subtracts the foil match values. The resulting difference distribution $d$ is also normal with mean $\mu_d$ and variance $\sigma_d^2$:

$$d \sim N\left(\mu_d, \sigma_d^2\right); \; \mu_d = k^T\mu; \; \sigma_d^2 = k^T\Sigma k.$$

In this example $\mu_d = 1 + w$ and $\sigma_d^2 = (8 + 8r)\sigma^2$. Then, the probability of selecting the old item is the probability that a sample from this difference distribution lies above zero, i.e. $q = 1 - \Phi(-\mu_d/\sigma_d)$.

If the foil is drawn from the other studied category, then the covariance matrix remains the same as when the foil is novel because the target and foil were still generated independently from one another. However, the match between the foil and the 2 studied items from the other category leads to $\mu = [1, w, 0, 0, w, w, 0, 0]^T$, so $\mu_d = 1 - w$ and $\sigma_d^2 = (8 + 8r)\sigma^2$. If, however, the foil is drawn from the same category as the old item, the mean is the same as if the foil is from a different category, but the covariance becomes

$$\Sigma = \begin{bmatrix} \sigma^2 & r\sigma^2 & 0 & 0 & r\sigma^2 & r\sigma^2 & 0 & 0 \\ r\sigma^2 & \sigma^2 & 0 & 0 & r\sigma^2 & r\sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & r\sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & r\sigma^2 & \sigma^2 & 0 & 0 & 0 & 0 \\ r\sigma^2 & r\sigma^2 & 0 & 0 & \sigma^2 & r\sigma^2 & 0 & 0 \\ r\sigma^2 & r\sigma^2 & 0 & 0 & r\sigma^2 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & r\sigma^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & r\sigma^2 & \sigma^2 \end{bmatrix}.$$

such that $\mu_d = 1 - w$ and $\sigma_d^2 = 8\sigma^2$. Although the mean difference is the same, the similarity between the target and foil reduces the variance such that more of the difference distribution falls above zero, leading to greater accuracy and an explanation for the Tulving effect (Tulving, 1981; Hintzman, 1988; Clark, 1997).

**Encoding strength** We allow items to vary in the strength with which they are encoded; a less strongly encoded trace will lead to a weaker match. Encoding strength may vary as a function of, for example, study time, but may also vary as a function of task structure, e.g., category size. To know whether such an effect exists, we assume that the exemplars of categories of different size may be encoded with varying strength. The encoding strength of a category, $c$, is a free variable. However, to avoid over-parameterization, we assume that singletons—items from categories of size 1—are encoded with strength 1 and only allow the strengths of items from larger category sizes to vary.

Encoding strength has a multiplicative effect on match strength. Thus, generalizing from the above examples, the mean match value for an old choice item is $\mu_O = s_O[1 + w(N_O - 1)]$, where $s_O$ is the encoding strength for the category from which the old item is drawn and $N_O$ is the number of items studied from the old category. Similarly, the mean match value for a new item is $\mu_N = s_N N_N w$, where $s_N$ and $N_N$ are the encoding strength and number of studied items for the category from which the new item is drawn. If the new item is novel (there were no similar items studied), then $N_N = 0$ and $\mu_N = 0$. The mean and variance of the difference distribution can then be expressed

$$\mu_d = \mu_O - \mu_N = s_O[1 + w(N_O - 1)] - s_N N_N w \tag{1}$$

$$\sigma_d^2 = 2\sigma^2 \left[ N + r \sum_{i=1}^{C} N_i(N_i - 1) - I(O = N) r N_O^2 \right] \tag{2}$$

where $C$ is the number of categories in the study list, $N_i$ is the number of studied exemplars from category $i$, and $I()$ is an indicator function that equals one when its argument—in this case, whether the old and new item are drawn from the same category—is true and is zero otherwise.

**Valence** All study items in both experiments were paired with a valence, although only in Experiment 1 were there negative valences. Thus, we re-frame the recognition task as selecting the item with the highest valence, rather than with the highest match value. Incorporating valence introduces other complications: 1) just as there is variability in the strength with which items are encoded, there is likely to be variability in the probability that the valence of an item is encoded; 2) differential attention to negative and positive items may lead to different encoding strengths depending on valence; and 3) positive and negative valences may be given different weight at the decision stage.

For each category $i$, we assume there is a probability of encoding its valence when it is positive, $p_i$, and when it is negative, $q_i$. If the valence of a category has not been en-

Table 1: Priors and posterior means and 95% HDI's for each parameter in the model. See main text for details.

| Exp. | Param. | Prior | Posterior mean (95% HDI) |
|---|---|---|---|
| 1 | t | G(0:001;0:001) | 0.096 (0.048–0.173) |
| | r | $G^{\frac{1}{2}}$(0:001;0:001) | 0.128 (0.072–0.208) |
| | w | B(1;1) | 0.355 (0.283–0.426) |
| | r | B(1;1) | 0.124 (0.013–0.288) |
| | $s_4$ | G(0:001;0:001) | 0.945 (0.784–1.123) |
| | $s_8$ | G(0:001;0:001) | 0.776 (0.632–0.927) |
| | l | G(0:001;0:001) | 3.478 (1.887–6.973) |
| | h | G(0:001;0:001) | 0.206 (0.086–0.348) |
| | $p_4$ | B(1;1) | 0.784 (0.711–0.855) |
| | $q_4$ | B(1;1) | 0.889 (0.670–0.999) |
| | $p_8$ | B(1;1) | 0.578 (0.482–0.660) |
| | $q_8$ | B(1;1) | 0.115 (0.003–0.322) |
| 2 | t | G(0:001;0:001) | 0.145 (0.061–0.286) |
| | r | $G^{\frac{1}{2}}$(0:001;0:001) | 0.140 (0.071–0.247) |
| | w | B(1;1) | 0.329 (0.284–0.378) |
| | r | B(1;1) | 0.038 (0.001–0.124) |
| | $s_2$ | G(0:001;0:001) | 0.879 (0.725–1.053) |
| | $s_{4I}$ | G(0:001;0:001) | 0.638 (0.518–0.774) |
| | $s_{4B}$ | G(0:001;0:001) | 0.516 (0.410–0.635) |
| | $s_6$ | G(0:001;0:001) | 0.493 (0.411–0.584) |

coded, we assume that the participant "guesses" that it is positive with probability $\frac{1}{2}$. This retrieved valence $v_i^0$ is used at decision instead of the true studied valence $v_i$. Regardless of whether the valence is retrieved correctly, if a category was assigned a negative valence at study, the encoding strength of the exemplars from that category is multiplied by a factor h, h > 0, which allows for negatively valenced items to be encoded with either greater or lower fidelity. Finally, at the decision stage, if the *retrieved* valence of an item is negative, its match is weighted by l, which can reflect loss aversion (l > 1) or risk-seeking (l < 1). Thus, the final expression for the mean of the difference distribution is

$$\mu_d = v_O^0 \mu_O h^{I(v_O<0)} l^{I(v_O^0<0)} \quad v_N^0 \mu_N h^{I(v_N<0)} l^{I(v_N^0<0)}. \qquad (3)$$

**Individual differences** For simplicity, we assume that individuals differ only in their encoding variability, i.e., $s^2$. The value of $s^2$ for a participant is assumed to be drawn from a group Gamma distribution parameterized by a mean t and standard deviation (shape $\frac{t^2}{r^2}$, rate $\frac{t}{r^2}$). All other parameters are assumed shared between participants.

## Parameter Estimation

To obtain parameter estimates, the model was implemented as a hierarchical Bayesian model in JAGS (Plummer, 2011). Given the predicted probability of choosing the old item ($q_i$) for each of the T total trials, the likelihood is Bernoulli: $\prod_{i=1}^{T} q_i^{y_i}(1 \quad q_i)^{(1 \quad y_i)}$, where $y_i = 1$ if the old item was chosen on trial i and is zero otherwise. Prior distributions were left vague. Posterior estimates are based on a sample of 5000 points from the posterior, after 1000 samples of burn-in.

# Model Fits

The model was fit to each experiment separately. The prior distributions and estimated posterior means and 95% Highest Density Intervals (HDI's) are given in Table 1.

## Experiment 1

Observed and predicted mean probabilities of choosing the old item are shown in Figure 3.

**Category size** As mentioned above, the encoding strength of a singleton was set equal to 1. The encoding strength for an item from a category with 8 exemplars ($s_8$) was credibly less than that of both a singleton (95% HDI for 1 $s_8$ = [0.08, 0.37]) and an item from a category with 4 exemplars ($s_4$; 95% HDI for $s_4$ $s_8$ = [0.06, 0.29])[1]. The encoding strength of a 4-item category was not credibly different from that of a singleton (95% HDI for $s_4$ 1 = [-0.22, 0.12]). Overall, then, items from categories with more exemplars tend not to be encoded as strongly. This could be a result of failure to encode distinctive features of items in favor of more holistic, prototype-like representations (Homa, Dunbar, & Nohre, 1992). It may also result from a threshold process in which only those memory traces that are sufficiently similar to a probe are activated and take part in the recognition process; if more traces are active, this introduces noise into the comparison process that can harm performance (e.g., Hintzman, 1988).

**Valence** Participants give credibly greater weight to (retrieved) negative values when deciding between two choice items (the 95% HDI for l is greater than 1), indicating that participants are loss-averse at the decision stage. Valence also has an impact on encoding: The encoding strength for an item assigned a negative value is credibly reduced relative to one with a positive one (95% HDI for h is less than 1). Further, the probability of correctly encoding the value increases when the positive and negative items are from two perceptually distinct 4-item categories, rather than from the same 8-item perceptual category (95% HDI for $p_4$ $p_8$ = [.11, .31]; 95% HDI for $q_4$ $q_8$ = [.55, .99]). Thus, although participants clearly want to avoid negative items, they encode the perceptual features of those items more poorly.

## Experiment 2

Observed and predicted mean probabilities of choosing the old item are shown in Figure 4.

**Category size** As in Experiment 1, categories with fewer studied exemplars tend to be encoded more strongly. Singletons are encoded more strongly than 6-item categories (95% HDI for 1 $s_6$ = [0.42, 0.59]) and 4-item categories both blocked (95% HDI for 1 $s_{4B}$ = [0.37, 0.59]) and interleaved (95% HDI for 1 $s_{4I}$ = [0.24, 0.49]), but not 2-item categories (95% HDI for 1 $s_2$ = [-.05, .28]). 2-item categories are encoded more strongly than 6-item categories (95% HDI for $s_2$ $s_6$ = [0.29, 0.48]), blocked 4-item categories (95% HDI for $s_2$ $s_{4B}$ = [0.23, 0.50]), and interleaved 4-item categories (95% HDI for $s_2$ $s_{4I}$ = [0.12, 0.38]). Finally, although interleaved 4-item categories are encoded more strongly than 6-item categories (95% HDI for $s_{4I}$ $s_6$ = [0.06, 0.24]), this is

[1]Two parameters are said to be credibly different if the 95% HDI of their posterior difference excludes zero.

not true for blocked 4-item categories (95% HDI for $s_{4B} - s_6$ = [-0.07, 0.10]).

**Blocked vs. interleaved** Interleaved presentation results in stronger encoding of the individual exemplars than does blocked presentation (95% HDI for $s_{4I} - s_{4B}$ = [0.02, 0.25]). This implies that a category size effect may not be due solely to the number of studied exemplars; after all, if a list contains more items from a category, those items are also more likely to be studied together if the study list is randomly ordered. It would appear that increased category size as well as blocked study may independently contribute to weaker encoding of exemplars, leading to a representation that is more "prototypical".

## Discussion

The more similar items are stored in memory, the more they tend to interfere with one another (as in the homogeneity effects of Kahana & Sekuler, 2002); conversely, the more distinctive an item is (e.g., a singleton), the stronger it is encoded. Interleaved presentation tends to counter these effects. This suggests an encoding process whereby, if the current study item is sufficiently similar to the preceding study item, attention is directed only to similar features, leading to weaker encoding of the individual items. It may also be that the two items end up being encoded in the same memory trace, rather than separate traces; this composite trace (e.g., Howard & Kahana, 2002) might itself be encoded relatively strongly, but does not store much of the individual variation in exemplars. When successive study items are dissimilar, individuating features are preserved either through stronger encoding of individual traces or the failure to "blend" the two items into a single composite trace.

Items are also stored less strongly when they are assigned a negative valence, even though participants demonstrate loss-aversion at the decision stage. Given this loss-aversion, participants may attend more to the negative value and thereby fail to encode the item's perceptual features. Increased attention to the negative value—and away from the item itself—may also result from the novelty/distinctiveness of the negative value; after all, negative values do not occur as often over the course of the experiment. It may also reflect an encoding strategy that results in effects analogous to those of blocked study, that is, participants may attempt to find and encode features that are *shared* among negative items, thus making them easier to detect on the basis of those features (e.g., "a spoke on the upper left" might indicate negativity). This strategy only works, of course, if the features shared by negative items are *not* shared by positive items; if all items come from the same perceptual category, this strategy would only lead to poor overall performance, as observed. In any event, our results contrast with findings of memory enhancement for negative stimuli (e.g., Kensinger & Corkin, 2003), although this is likely due to the fact that the valence is not inherent to our stimuli, but is assigned arbitrarily.

In this paper, we demonstrated how well-controlled artificial stimuli and a reasonably open-ended model can be used to jointly investigate a variety of memory phenomena in a reasonably "pure" setting, with minimal preexperimental associations or strategies. A fruitful avenue of future research would be to vary between-category similarity in order to discover when items become "sufficiently" similar to lead to the observed blocked/interleaved effect. Varying the magnitude of values, rather than just valence, will provide more information about induced strategic encoding effects. In addition, an entire motion trajectory was obtained on each trial of the present experiments; future analysis of this data will yield even more insight than the simple choice behavior reported here.

## References


Carvalho, P. F., & Goldstone, R. L. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. *Proc. of the 34th Annual Conference of the Cognitive Science Society*, 186–191.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 232–238.

Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Homa, D., Dunbar, S., & Nohre, L. (1992). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 444–458.

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

Kachergis, G., Recchia, G., & Shiffrin, R. M. (2011). Adaptive magnitude and valence biases in a dynamic memory task. *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, 819–824.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, *42*, 2177–2192.

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169–1180.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Plummer, M. (2011). *JAGS: Just another gibbs sampler*. Available from http://mcmc-jags.sourceforge.net/

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Shepard, R. N., & Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology*, *4*(3), 351–377.

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*(2), 267–287.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 479–496.