

# Recognizing spatial patterns: A noisy exemplar approach

Michael J. Kahana and Robert Sekuler  
Brandeis University

## In Press: Vision Research

Models of categorization typically rely on the use of stimuli composed of well-defined dimensions (e.g., F. G. Ashby and W. T. Maddox, 1993). We apply a similar approach to the analysis of recognition memory. Using a version of S. Sternberg's (1966) short-term recognition paradigm, we asked whether NEMO, a noisy exemplar summed-similarity model, could account for variation in mean performance on individual trials. NEMO provided a very good overall fit to recognition data from three experiments. However, its failure to fit data for certain lists of stimuli suggested a revision of the summed-similarity assumption. Our model-based analysis showed that subjects used inter-item similarity, in addition to probe-item similarity, as the basis for their decisions. This represents a major departure from existing recognition models that assume subjects' judgments depend exclusively on the summed similarity of the probe to the study items.

**Keywords:** Discrimination, Recognition, Multidimensional Scaling, Spatial Vision.

The last 20 years have produced powerful models of multidimensional identification and categorization (Ashby & Maddox, 1998; Estes, 1994; Maddox & Ashby, 1996; McKinley & Nosofsky, 1996; Nosofsky, 1986) and episodic memory (Chappell & Humphreys, 1994; Humphreys, Pike, Bain, & Tehan, 1989; Humphreys, Bain, & Pike, 1989; Masson, 1995; Murdock, 1997). This development has been paralleled by the creation of quantitative, physiologically-inspired accounts of visual detection and discrimination (Graham, 1989; Wilson & Wilkinson, 1997). Identification, categorization and memory are integral to performance on visual tasks; conversely, visual limitations can influence measures of identification, categorization and memory. Despite this clear connection between these domains, though, work in each has tended to proceed largely in isolation from the other.

---

The authors acknowledge support from National Institutes of Health grant MH55687. Some of this work was presented at the 1999 Meeting of the Association for Research in Vision and Ophthalmology. The authors thank Dan Rizzuto for introducing us to the use of genetic algorithms for non-linear optimization and for programming assistance. We also thank Arye Elfenbein for assistance with data collection. Correspondence concerning this article may be addressed to either Michael Kahana or Robert Sekuler, Volen National Center for Complex Systems, MS 013, Brandeis University, Waltham, MA 02254-9110. E-mail may be sent to kahana@brandeis.edu or sekuler@brandeis.edu

Categorization models represent items as points in complex multidimensional stimulus spaces (Ashby & Perrin, 1988; Maddox & Ashby, 1996; Nosofsky, 1992), with decision rules that can predict performance in a variety of classification paradigms (Maddox & Ashby, 1996; Nosofsky & Palmeri, 1998; Nosofsky & Alfonso-Reese, 1999). Although models of classification and models of visual discrimination share many assumptions about stimulus representation and subjects' decision rules, models of classification have been primarily developed to explain subjects' classification of combinations of simple geometric forms, whereas models of discrimination have been developed to explain subjects' discrimination of elemental visual stimuli, including sinusoidal luminance gratings. Because such stimuli can be combined to synthesize more complex images such as textures and natural scenes, they represent a natural test-bed for assessing theories' power and generalizability.

Many models of episodic memory share assumptions with discrimination and classification models. For example, most episodic memory models represent items as vectors in multidimensional feature space (e.g., Chappell & Humphreys, 1994; Howard & Kahana, 2002; Murdock, 1997; Shiffrin & Steyvers, 1997), and employ decision processes similar to those in many discrimination and classification models (e.g., Hockley & Murdock, 1987). Episodic memory models, however, have failed to ground their abstract stimulus representations in perceptually-defined similarity structures. This has undermined the models' ability to account for generalization, stimulus-dependent interactions, and other similarity-related

phenomena (Edelman, 1999). Finally, vision research, which has historically ignored memory, recently began to recognize memory’s important role in even simple perceptual tasks such as detection and discrimination (Blake, Cepeda, & Hiris, 1997; Kahana & Bennett, 1994; Magnussen, 2000; Magnussen, Greenlee, Asplund, & Dyrnes, 1991).

### *Elemental Visual Stimuli*

This paper examines episodic recognition of two-dimensional textures created by linearly summing trios of sinusoidal gratings. Our use of textures composed of simple gratings to study visual memory reflects well-established ideas about visual coding in the primate brain. Early in the processing stream that leads ultimately to vision, neurons in the mammalian primary visual cortex act as bandpass filters that extract retinal image structure on different spatial scales. The cells’ receptive fields vary along several dimensions, including size, spatial location, and orientation. Such neurons feed subsequent cortical regions reliable, detailed information about the size, location, and orientation of local features in the retinal image (Geisler & Albrecht, 1997).

We had several principal reasons for probing visual memory with compound sinusoidal gratings. First, at early stages of visual processing, elemental features such as spatial frequency, orientation, and components’ relative positions are extracted and explicitly represented in the activity within ensembles of neurons (De Valois & De Valois, 1988; Olzak & Thomas, 1999). Second, these same dimensions may form a basis of organization in short-term visual memory (Kosslyn et al., 1999; Magnussen, 2000; Magnussen & Greenlee, 1999). As a result, these stimuli allow us to control, isolate, or combine elemental features in order to test ideas about encoding and storage for memory. Third, unlike semantically-rich stimuli, such as words or images of meaningful objects, our stimuli are not burdened by complexities of extra-laboratory associations and thus resist symbolic coding. Fourth, simple, well-defined visual stimuli make it easy to manipulate similarity relations among stimuli because they have natural metric representations in multidimensional space. Because interstimulus similarities are at the heart of many theories of cognition, ability to manipulate similarity relations represents an important advance in testing and extending such theories (Teller & Palmer, 1998). Finally, with various simple visual stimuli, including gratings, successive discrimination is largely unaffected by a delay between study and test stimuli, at least for delays ranging up to a minute or more, unless some interfering stimulus intervenes (Blake et al., 1997; Magnussen, Greenlee, Asplund, & Dyrnes, 1990; Magnussen et al., 1991; Magnussen, Greenlee, & Thomas, 1996). This preservation of encoded information simplifies the design and interpretation of memory studies.

### *Discrimination vs. Recognition*

Numerous studies of visual discrimination (Graham, 1989) have used stimuli like the ones used in our recognition memory experiments. So it is worthwhile to consider the

relation between discrimination and recognition tasks, that similarities between the two tasks can be understood more clearly.

In a typical discrimination task, subjects view two stimuli,  $s_1$  and  $s_2$ , and attempt to order them according to some fixed visual criterion. For example, a participant may judge  $s_1$  and  $s_2$  as “same” or “different.” Note that this judgment can be made solely on the basis of the two stimuli presented on that trial. Two different strategies are used to aggregate data. Accuracy can be assessed for various pairs of stimuli, and results expressed as percent correct; or, accuracy can be expressed in a stimulus metric, e.g. as a threshold or critical stimulus value. Because discrimination usually focuses on sensory rather than mnemonic factors, the interval between  $s_1$  and  $s_2$  is kept short.

In an episodic recognition task, subjects view a list,  $S$ , of two or more stimuli. In the “yes”–“no” procedure, the participant sees a probe stimulus,  $p$ , and judges whether or not it occurred in  $S$  (in most studies, probes are as likely to occur in the list as not). Good performance on any trial cannot be achieved solely on the basis of visual information; the visual response evoked by the probe stimulus must be compared to information held in memory, i.e. some stored, internal representation of  $S$ .

Discrimination and recognition both require sensitivity to inter-stimulus similarities and differences. While discrimination performance directly reflects the similarity between  $s_1$  and  $s_2$ , recognition requires subjects to assess the similarity of the probe,  $p$ , to each of the items in the list,  $S$ .

### *Stimuli and Experiments*

We describe two studies of recognition memory for 2-D textures. In both experiments, stimuli are textures whose luminance profiles  $L(x, y)$  are:

$$L_{avg} [1 + A_1 \cos(\pi f x) + A_2 \cos(2\pi f x + \phi) + A_3 \cos(\pi g y)]$$

where  $L_{avg}$  is mean luminance;  $f$  is the spatial frequency of the vertical fundamental component, in cycles per degree;  $g$  is the frequency of the horizontal component;  $A_1$ ,  $A_2$ , and  $A_3$  are each component’s contrast; and  $\phi$  is the relative spatial phase of harmonic to fundamental. (Relative phase is components’ spatial positions relative to one another.) The two vertical components had a constant, harmonic relationship of 1 : 2, with contrasts in a 2 : 1 relationship. All contrast components were well above detection threshold; overall contrast was 0.4. Stimuli were 5 degrees visual angle in diameter, windowed with a 2 –  $D$  Gaussian function, and were viewed centrally. Stimuli were generated by Matlab software, including Matlab’s Image Processing Toolbox, and routines in the Psychophysical Toolbox (Branard, 1997) To avoid spatial non-linear distortions, display luminances were linearized using calibration routines from Pelli’s (1997) VideoToolbox.

For our first experiment, we created lists of 1–4 textures by randomly sampling textures from a pool of 27 textures generated by varying the three dimensions across three levels. Because sampling was without replacement, no texture appeared more than once on a trial. This experiment assessed

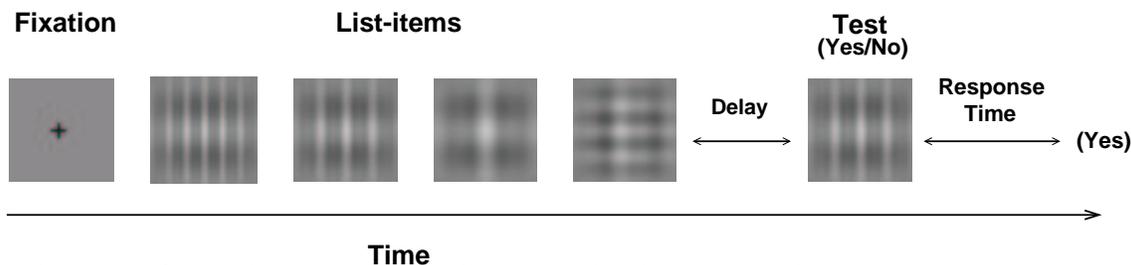


Figure 1. Trial structure for Experiment 1. This figure illustrates a trial with 4 textures (trials had 1, 2, 3 or 4 textures randomly drawn from a pool of 27 stimuli, as described in the text).

the effects of list length and recency on subjects' response accuracy. We conducted a second experiment with lists of only two stimuli each. Rather than allowing the stimuli to vary haphazardly along the three dimensions, we constrained the stimulus positions to either lie along a single dimension or to vary along all three dimensions simultaneously. This second experiment also parametrically varied the distances between  $s_1$  and  $s_2$  and  $p$ . Both experiments afforded stimulus lists and conditions that were likely to generate a wide range of recognition accuracy. This, we expected, would provide the strongest test of our theoretical account of the results.

## Experiment 1

### Subjects

Subjects were nine undergraduate and graduate students who participated for payment. Each participant gave 1800 trials across 5 sessions.

### Methods

Twenty-seven stimuli were generated from factorial combinations of three values each of  $f$ ,  $g$ , and  $\phi$ :  $f = 0.4, 0.8$  and  $1.2$  cycles/degree;  $g = 0.4, 0.8$ , and  $1.2$  cycles/degree;  $\phi = 0, \pi/4$ , and  $\pi/2$ . All frequency components were highly visible (overall contrast = 0.4) and had a mean luminance of  $18.7 \text{ cd/m}^2$ . All subjects had normal contrast sensitivity as assessed by the Pelli-Robson chart (Pelli, Robson, & Wilkins, 1988).

On each trial, subjects saw a list of 1, 2, 3, or 4 textures randomly chosen from the pool of 27 textures. Figure 1 shows the trial structure and sample stimuli. After the last stimulus of any list, a warning tone sounded and then a probe stimulus appeared for 1 s. Subjects judged whether the probe had appeared in the list. Half the time it had appeared, half the time it had not. Accuracy and latencies were recorded, and feedback about response correctness was given after each trial.

### Results

We separately analyzed trials on which the probe had been in the list and trials on which the probe item was a lure, that is was not in the list. As shown in Figure 2, mean hit-rate

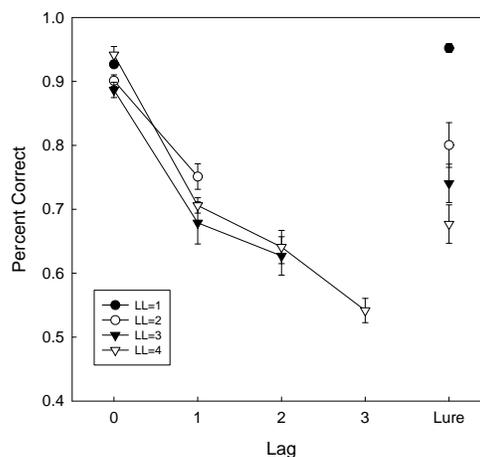


Figure 2. Accuracy as a function of recency (number of stimuli intervening between study and probe presentation) for lists of 1, 2, 3 and 4 items, and for lures (shaded area to the right of the graph). Error bars represent  $\pm 1$  standard error of the mean

(correctly identifying a probe as a member of the list) declined dramatically with the number of items intervening between study and test (lag). For a given lag, the number of items preceding a target did not greatly affect the hit rate. As shown by the points on the right side of the figure, subjects' ability to recognize that a test lure had not appeared in the list declined sharply as list length increased. These effects were also seen in RT data. Across conditions, the correlation between accuracy and RT was  $-0.95$ .

Our probe recognition data exhibit a much larger recency effect than is seen in probe recognition studies using rehearsable, verbal materials (e.g., Sternberg, 1975; Kahana & Loftus, 1999). In fact, in our study, the list-length effect may be entirely a consequence of recency (e.g., memories decaying with time or the presence of intervening events). Alternatively, it could be that the effect is due, in part, to the increasing probability that as a list grows, so does the chance that it contains at least one item that is perceptually similar to the lure.

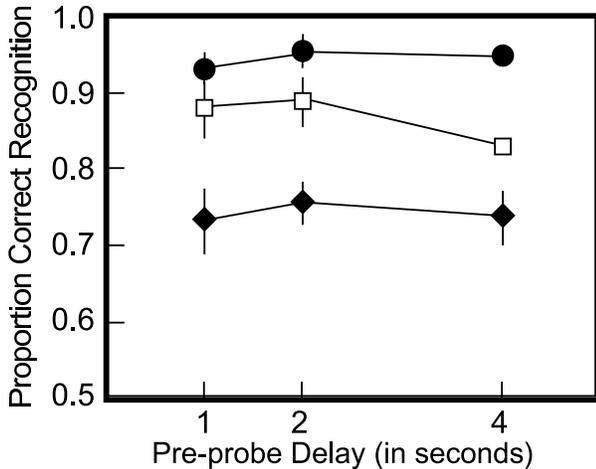


Figure 3. The effect of an unfilled delay on recognition accuracy. The three curves represent different list lengths (circles for LL=1, squares for LL=2, and diamonds for LL=4). Error bars represent  $\pm 1$  standard error of the mean.

To determine whether the large recency effects seen in Figure 2 were due to the passage of time, we conducted a subsidiary experiment, varying the delay between the presentation of the last item in the study list and the presentation of the probe. Each of five subjects generated 945 responses, over three sessions. Delays were either 1 s (as in Experiment 1), 2 s, or 4 s. Figure 3 shows that recognition accuracy was essentially unchanged by delay duration. We take two lessons from delay’s failure to influence performance: first, in the absence of interfering events the most recently presented item is unaffected by delay; second, that the memory representation of items prior to the most recent one is similarly unaffected. So, in this case, discrimination (list length=1) and recognition (list length > 1) work the same way.

As in previous recognition memory studies, the analysis of our first experiment averaged data from many different individual lists. For example, the data shown in Figure 2 for lists with two study items is an average across 48 different lists of stimuli. As is normative for studies of recognition memory, lists of study items and probe were generated at random. As a result, the data for any given list length represent mean performance calculated for lists comprising different study-test items. The process of aggregating diverse lists masks information potentially quite valuable for understanding memory. Specifically, with any given list length, some lists contained study items that were relatively similar to one another, but other lists’ items were less similar to one another. And the same point holds for the diverse, pairwise similarity relations between each study item,  $s_1$  or  $s_2$ , and the probe,  $p$ . Although it is convenient to work with the average performance for all exemplars of a given list length, intuition suggests that the diversity of the individual exemplars is probably matched by diversity in the performance produced by various lists.

Stimuli, like the ones used in our first experiment, are well

suited for taking the behavioral analysis one level deeper than usual — looking not just at mean accuracy across conditions, but also examining performance with individual stimulus lists. To facilitate this treatment of the data and to guide subsequent empirical work we developed a multidimensional, signal detection model of episodic recognition, and applied it to accuracy data on each of the individual trials with the 355 unique lists in Experiment 1.

## NEMO

For convenience we refer to our model as NEMO (Noisy-exemplar Model). NEMO represents each stimulus as a multivariate normal distribution in feature space (e.g., Ashby & Maddox, 1998; Ennis, Palen, & Mullen, 1988). These distributions arise from several sources of external and internal noise (Pelli & Farell, 1999). Borrowing ideas from exemplar models of classification (McKinley & Nosofsky, 1996; Nosofsky, 1986) we assume that as each stimulus is presented its noisy coordinates are stored in memory, and that judgments are based, in large part, on the summed similarity between the probe item and these stored representations. In particular, summed similarity refers to the sum of pairwise similarity measures between the probe, on one hand, and the representations of each of the study items, on the other. Borrowing ideas from decision-bound models of human classification (Ashby & Maddox, 1998), NEMO uses a deterministic response rule, responding “yes” if the summed similarity crosses a decision bound (criterion) that separates targets and lures.

*Summed Similarity and Optimal Criterion:* In NEMO, the basic computation underlying recognition performance is the summed pairwise similarity between each stimulus item’s noisy representation and the relatively noiseless representation of the probe. If stimuli are randomly selected from a multidimensional space, the summed similarity of a target to the contents of memory will typically exceed the summed similarity for a lure. This provides a basis for modeling two-alternative forced-choice recognition. For “yes”–“no” recognition, as in the Sternberg procedure, prior experience must be used to determine a threshold value that distinguishes between targets and lures.

We assume that the participant uses an optimal decision criterion to decide whether the summed similarity was more likely to have come from the presentation of a lure or from the reappearance, as a probe item, of a stimulus that was in the list. Presumably, experience enables participants to adjust their criteria so as to suit the context (e.g. Treisman & Williams, 1984). Because summed similarity increases with list length, we determine the optimal criteria separately for each list length.

*Similarity and Distance.* Following exemplar-based classification models, we define the similarity,  $\eta_{s_i, s_j}$  among two representations,  $s_i$  and  $s_j$ , as given by:

$$\eta(s_i, s_j) = e^{-\tau d(s_i, s_j)^c}, \quad (1)$$

where  $d$  is the weighted distance between the two stimulus vectors, and  $\tau$  and  $c$  jointly determine the form of the gen-

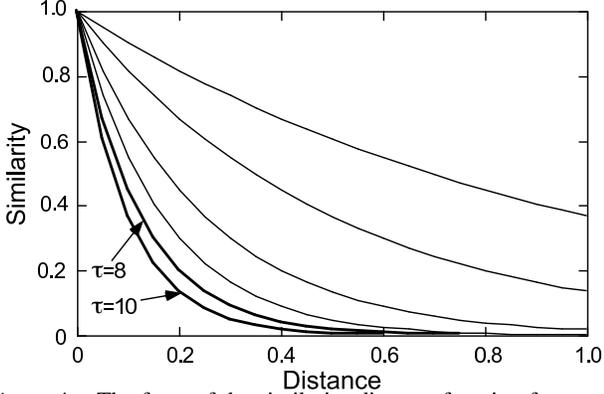


Figure 4. The form of the similarity-distance function for  $c = 1$ , and  $\tau$  varying from 1 (top curve) to 10 (bottom curve). The bold curves at  $\tau = 8$  and  $\tau = 10$  depict the best fitting function used in fitting data from Experiments 1 and 2 respectively.

eralization gradient. When  $c = 1$ , the model implements a simple exponential generalization function; when  $c \geq 2$  the model implements a sigmoidal generalization function. As shown in Figure 4, increasing the value of  $\tau$  causes similarity to decrease more rapidly with increasing distance. The distance along each dimension,  $s_i(k) - s_j(k)$ , is weighted by a factor  $w_k$ , to ensure (i) that the measurement is not sensitive to absolute variations in the scale of the dimensions (e.g., spatial frequency in cycles/degree vs. phase in radians), and (ii) that the model can capture global differences in the attention that each dimension attracts. For example, subjects could divert attention from one of the dimensions because that dimension is relatively more difficult to encode reliably, and/or because it produces too much interference during retrieval. Although it is likely that such attentional strategies would change as a function of practice, for simplicity we assume that the weighting dimensions remain constant over the course of the entire experiment. In combining the distances along each dimension we adopt a simple Euclidean metric. In preliminary applications of NEMO to our data we used a generalized Minkowski distance, and found that the best fitting value of the Minkowski exponent remained close to 2. This suggests that Euclidean distance underlies the representational (psychological) similarity among our compound stimuli.

*Noise.* As described above, we assume that each stimulus is stored imperfectly in memory. To the stored representation of each item we add a noise vector,  $\epsilon$ , whose components are mean-zero Gaussians whose variance depends on the stimulus dimensions comprising that item, and on the recency of that item’s occurrence. In using this noise term, we can adopt a fully deterministic decision rule (e.g., Ashby & Maddox, 1998). Variability in subjects’ responses from one occurrence of an item to another are modeled by the sampling of each item from its noisy representation. To simulate forgetting, we assume that the most recent stimulus contributes the most to the summed similarity and that earlier items contribute less and less. Stimulus-dependent noise varies across

dimensions, but the rate of forgetting does not. Figure 5 schematizes the application of NEMO to a trial consisting of three study stimuli followed by a “lure”.

*Summed similarity and inter-stimulus similarity:* NEMO departs from traditional summed similarity models by assuming that subjects’ responses are determined not only by the similarity between list items and the probe, but also by the similarities among list items themselves. This introduction of interstimulus similarity, coupled with the noisy coding of exemplars, distinguish this model from most existing exemplar models of categorization and recognition.

Before introducing the details of interstimulus similarity, we present the decision rule for the model based entirely on summed similarity of list items to the probe. Given a list of items,  $\mathbf{s}_1 \dots \mathbf{s}_L$ , and a probe item,  $\mathbf{p}$ , a summed similarity version of NEMO will respond “yes” if:

$$\sum_{i=1}^L \alpha_i \eta(\mathbf{p}, \mathbf{s}_i + \epsilon) > C_L \quad (2)$$

where  $\epsilon$  is a vector representing the noise associated with each stimulus dimension, and  $C_L$  represents the optimal criterion for a list of  $L$  items. In this formulation only the distances between the list items and the probe enter the decision rule.

Consider the predictions this model makes for a list of two items,  $\mathbf{s}_1$  and  $\mathbf{s}_2$ , followed by a probe item,  $\mathbf{p}$ , that is very similar to  $\mathbf{s}_1$ . If  $\mathbf{s}_2$  is also similar to  $\mathbf{p}$ , summed similarity will increase, making it more likely that the subject will make say “yes”. Moving  $\mathbf{s}_2$  far away from  $\mathbf{s}_1$  should increase the likelihood that the subject will correctly reject the lure.

One could imagine, however, that subjects are sensitive to the configuration of stimuli, that is, the layout and ordering in perceptual space of successive stimuli (Blaser, Pylyshyn, & Holcombe, 2000; Leopold, O’Toole, Vetter, & Blanz, 2001). In the example above, one might conjecture that if  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are very far apart, subjects will perceptually group the lure with  $\mathbf{s}_1$ . In this case, subjects will be *more* likely to make a false alarm.

To allow for the possibility that subjects incorporate interitem similarity into their decision rule, we add together the summed similarity and the interstimulus similarity, weighted by a parameter  $\beta$ . If  $\beta = 0$  the model reduces to the summed similarity model given in Equation 2. If  $\beta$  is negative, a given lure will be more tempting when  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are widely separated. Conversely, if  $\beta$  is positive, a lure will be less tempting when  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are widely separated.

Given a list of items,  $\mathbf{s}_1 \dots \mathbf{s}_L$ , and a probe item,  $\mathbf{p}$ , NEMO will respond “yes” if:

$$\sum_{i=1}^L \alpha_i \eta(\mathbf{p}, \mathbf{s}_i + \epsilon) + \frac{1}{L(L-1)} \beta \sum_{i=1}^{L-1} \sum_{j=i}^L \eta(\mathbf{s}_i + \epsilon, \mathbf{s}_j + \epsilon) > C_L \quad (3)$$

*Simplifying assumptions:* In applying NEMO to our data, we made a number of simplifying assumptions (listed below). A proper evaluation of each of these assumptions would require experiments designed specifically for that purpose. We leave this for future work.

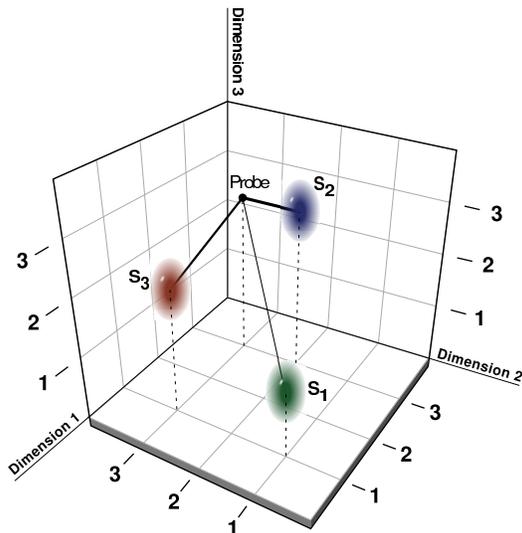


Figure 5. Cartoon of NEMO applied to a 3-item recognition task. Stimuli, varying along three dimensions, are depicted by fuzzy circles. The shaded ellipses represent the amount of noise associated with the coding of the stimulus. The black circle indicates the probe item. In this example, the probe is a “lure”, that is, it was not present in the study list. Recognition judgments are based on the summed similarity between the probe and the studied items. Similarity decreases monotonically with increasing distance, as shown in Figure 4. The functional relation between similarity and distance is given in Equation 1.

- Independence of noise along each of the stimulus dimensions (i.e., the covariances associated with  $\epsilon$  were set to zero).
- Equal rates of forgetting along each of the three stimulus dimensions. This is because the forgetting parameter,  $\alpha$ , is a single scaling parameter applied to the similarity between  $\mathbf{p}$  and each of the study items.
- Rationality. NEMO chose an optimal decision criterion, halfway between the means of the old and new item distributions (to do this we ran the simulation twice for each participant, once to determine the distributions of old and new items and then a second time to generate the predicted values for each trial).
- Euclidean Distance. After preliminary simulation results suggested that a Euclidean distance provided an equal or better fit than other Minkowski exponents, we used Euclidean distance in all subsequent runs of the model.

Even with these reasonable simplifying assumptions, the full implementation of NEMO included 11 free parameters for Experiment 1 and 9 free parameters for Experiment 2. After presenting the models fits to data from each experiment we will discuss the consequences of relaxing some of these assumptions.

## Modeling Experiment 1

We fit NEMO to participants’ accuracy for each of the 355 different lists by minimizing the root-mean-squared-difference (RMSD) between observed and predicted values.<sup>1</sup> Because NEMO implements a forgetting process, it captures the basic serial position effects in the data. Across the 14 conditions representing different list lengths and item types (targets from serial positions 1– $L$ , and lures), the correlation between predicted and average observed values was 0.92.

As noted earlier, each list consisted of randomly chosen stimuli with varying similarity relations, both between the probe the stimuli, and among the stimuli themselves. A model such as NEMO allows us to take our behavioral analysis one level deeper, examining NEMO’s fit to mean accuracy for each of our randomly-generated (but unique) lists of study and probe items.

Figure 6 shows that NEMO accounts for a significant portion of the variability across lists. Although the variance accounted for was not spectacular ( $r=0.80$ ), it should be noted that the maximal correlation is greatly limited by the variability associated with individual lists. That is, with only five repetitions of each list per subject there is considerable uncertainty over the estimated mean accuracy for each of the 355 lists.

Table 1 gives the best fitting model parameters derived from the genetic algorithm. The first three parameters are the variances of the horizontal frequency, vertical frequency, and the phase of the second vertical component (we assumed that the covariance terms were zero). These parameters characterize the variance-covariance matrix of  $\epsilon$  in Equation 2. The next three parameters model forgetting. In computing the sum of the similarity between the probe and each of the list items, the contributions of recent items are weighted more

<sup>1</sup> Preliminary attempts at fitting NEMO to the experimental results revealed that the error surface contained many local minima. Standard optimization techniques like downhill SIMPLEX (Nelder & Mead, 1965) encounter difficulties with such error surfaces, which makes such techniques suboptimal for these problems. Methods based on genetic algorithms (Mitchell, 1996) are especially appropriate when dealing with such complex optimization problems. In particular, genetic algorithms produce a global minimization, mitigating the potential pitfalls of local minima. We therefore used a genetic algorithm, varying 9 parameters to minimize the RMSD between observed and predicted values. A population of 1000 random parameter sets (“individuals”) was evolved for 20 generations. After each generation, an individual’s fitness was defined by the negative of its RMSD value. Each individual was run for 300 trials on each list. By using this large number of trials (as compared with 45 in the experimental data — each of 9 subjects got each list 5 times, once per session), we reduced noise associated with modeling error. At the end of every generation, each of the 500 least-fit individuals was replaced with new individuals. These new individuals were generated from the two best fitting individuals (parents) by randomly drawing each of their parameter values from one of their parents. The 500 individuals with the best fitness were mutated by a single, Gaussian parameter change with a standard deviation of 30% of the parameter’s range. Using this technique, the best fitting parameter values rarely improved after the 10th generation.

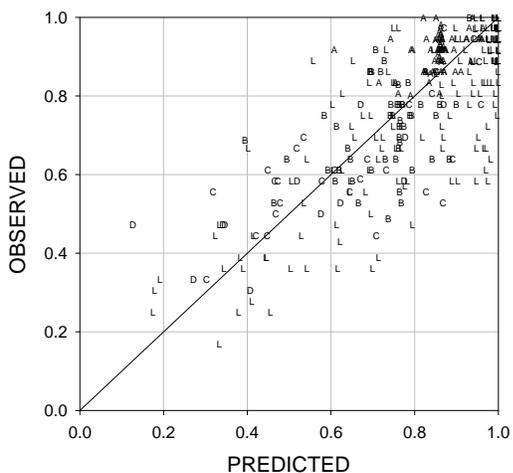


Figure 6. NEMO’s attempt to fit accuracy data from individual lists in Experiment 1. Observed and predicted results are plotted against one another. The symbols A–D denote targets, with 0–3 items intervening between study and test; L denotes a lure trial.

heavily than earlier list items. The most recent item has weight of 1, the next most recent has weight  $\alpha_1$ , the third most recent has weight  $\alpha_2$ , and the fourth most recent item has weight  $\alpha_3$ . The last two parameters,  $\tau$  and  $c$ , determine the tuning of the similarity function (see Equation 1).

After correcting for differences in the weighting parameters of the three dimensions, the noise associated with storing phase information was about 50% larger than that associated with storing spatial frequency information. The values of the  $\alpha$  parameters show that NEMO weights older items less heavily in its decision process, with the earlier list item contributing less than half of the similarity of the most recent list item.

One of the most interesting findings of the modeling exercise was that  $\beta$  took on such a large negative value. This value of -0.84, indicates that mean interitem similarity is as important in determining subjects’ responses as the similarity of the probe to a given list item. In the case of a two-item list the model is asking whether the (weighted) mean similarity of  $\mathbf{p}$  to  $\mathbf{s}_1$  and  $\mathbf{s}_2$  greater than  $\mathbf{s}_1$ ’s similarity to  $\mathbf{s}_2$  (plus a constant).

To determine whether the  $\beta$  parameter was actually responsible for a significant improvement in goodness-of-fit, we simulated a restricted version of NEMO with  $\beta = 0$ . Because the models are nested we were able to conduct a  $\chi^2$  test for the difference in the goodness-of-fit of the two models. The generalized model yielded a significant improvement in goodness-of-fit ( $\Delta(df) = 1, \Delta\chi^2 = 10.3, p < 0.005$ ), indicating that the  $\beta$  parameter is playing an important role in the success of the model.

Table 1  
Best fitting parameter values for NEMO’s fit to the data from Experiment 1

| Parameter         | Value |
|-------------------|-------|
| $\sigma_{f_v}$    | 0.083 |
| $\sigma_{f_h}$    | 0.071 |
| $\sigma_{\phi_v}$ | 8.35  |
| $w_v$             | 1.0   |
| $w_h$             | 1.16  |
| $w_{\phi_v}$      | 0.012 |
| $\alpha_1$        | 0.80  |
| $\alpha_2$        | 0.74  |
| $\alpha_3$        | 0.47  |
| $\beta$           | -0.84 |
| $\tau$            | 8.8   |
| $c$               | 1.08  |
| RMSD              | 0.120 |

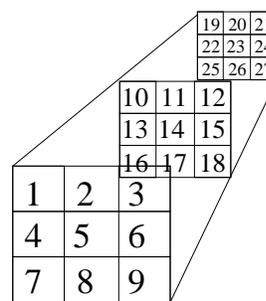


Figure 7. Codes corresponding to the 27 stimuli used in Experiment 1. The three matrices represent different levels of vertical frequency (1-9 being lowest and 19-27 being highest). Within each matrix, horizontal frequency increases from the top row to the bottom row and phase increases from the left column to the right column.

To better understand the variability that NEMO failed to account for, we identified lists that produced significant discrepancies between observed and predicted values. Table 2 shows lists where deviation between model and data was statistically significant at the  $p < 0.001$  level (Bonferroni corrected). Because we collected a fair amount of data on each list, we adopted this stringent criterion for classifying NEMO’s failures. Out of our 355 conditions, 23 met this criteria. The first two columns give the percent correct as predicted by NEMO, and as observed in our data, respectively. The third column indicates the serial position of  $\mathbf{p}$ , if it was on the list; otherwise, it notes that  $\mathbf{p}$  was a lure. The last five columns represent the actual stimuli that were presented (the probe followed by the 1–4 list items). These numeric codes correspond to the positions in a three dimensional space, as shown in Figure 7.

Most of the lists that stumped NEMO probed memory with a lure (17/23). On more than 3/4 of these lists, NEMO predicted higher performance than seen in the data. The random nature of the stimulus-probe arrangements made it difficult to see any clear patterns in NEMO’s failures. This led us to

Table 2

*NEMO's failures to fit accuracy data from Experiment 1*

| NEMO | Data  | Serial Position | $\mathbf{p}$ | $\mathbf{s}_1$ | $\mathbf{s}_2$ | $\mathbf{s}_3$ | $\mathbf{s}_4$ |
|------|-------|-----------------|--------------|----------------|----------------|----------------|----------------|
| 0.89 | 0.639 | 1               | 12           | 12             | 5              | 21             |                |
| 0.84 | 0.639 | 2               | 15           | 11             | 15             | 13             |                |
| 0.83 | 0.944 | 2               | 25           | 23             | 25             |                |                |
| 0.61 | 0.917 | 3               | 18           | 11             | 12             | 18             |                |
| 0.74 | 0.944 | 4               | 19           | 6              | 15             | 3              | 19             |
| 0.85 | 0.999 | 4               | 19           | 7              | 21             | 9              | 19             |
| 1.00 | 0.889 | Lure            | 10           | 24             |                |                |                |
| 1.00 | 0.889 | Lure            | 22           | 7              |                |                |                |
| 0.76 | 0.972 | Lure            | 13           | 15             |                |                |                |
| 0.75 | 0.972 | Lure            | 9            | 5              |                |                |                |
| 0.32 | 0.444 | Lure            | 1            | 22             | 2              |                |                |
| 0.33 | 0.167 | Lure            | 8            | 21             | 7              |                |                |
| 0.99 | 0.833 | Lure            | 17           | 12             | 4              |                |                |
| 0.80 | 0.472 | Lure            | 27           | 17             | 2              |                |                |
| 0.99 | 0.889 | Lure            | 3            | 6              | 15             |                |                |
| 0.56 | 0.889 | Lure            | 19           | 5              | 10             |                |                |
| 0.98 | 0.722 | Lure            | 12           | 9              | 19             | 16             |                |
| 0.96 | 0.806 | Lure            | 14           | 7              | 3              | 11             |                |
| 0.87 | 0.667 | Lure            | 26           | 8              | 6              | 16             |                |
| 0.98 | 0.639 | Lure            | 21           | 10             | 8              | 25             |                |
| 0.90 | 0.583 | Lure            | 24           | 22             | 7              | 16             | 12             |
| 0.95 | 0.694 | Lure            | 26           | 11             | 3              | 18             | 10             |
| 0.71 | 0.389 | Lure            | 25           | 27             | 7              | 16             | 8              |

move away from randomly generated lists. In Experiment 2 we constructed specially designed lists to help us to see what conditions present the greatest challenges to NEMO.

## Experiment 2

### Subjects

Subjects were eight undergraduate and graduate students who participated for payment. Each participant gave 1500 trials across five sessions. All had normal contrast sensitivity as measured with Pelli-Robson charts.

### Methods

Although NEMO accounted for a fair portion of the variance in performance on individual lists in Experiment 1 (see Figure 2), the random selection of stimuli in that experiment limited our ability to examine effects of inter-stimulus similarity because that variable was not specifically controlled. We therefore designed and used sets of stimuli that embodied various patterns of inter-stimulus similarity. To simplify experimental design and to limit the number of trials, we used only lists comprising two stimuli,  $\mathbf{s}_1$  and  $\mathbf{s}_2$ . Each list was followed by a probe item,  $\mathbf{p}$  and as before, the probe was equally often a lure, or one of the study items. Stimuli were drawn from a 3-D stimulus space based on the same dimensions as Experiment 1, but allowing a somewhat wider variation along the phase dimension. NEMO's estimates of noise on the three stimulus dimensions were used to provide a rough scale of stimulus distances, transforming such distances into a similarity metric.

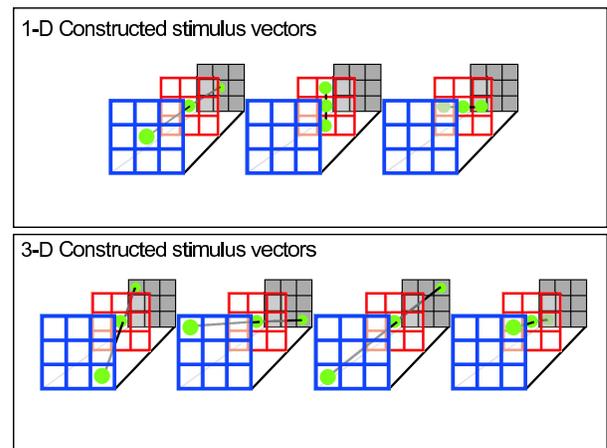


Figure 8. Illustration of the seven constructed stimulus vector types used in Experiment 2

For Experiment 2, stimuli and probes for each trial were constrained to lie along a single vector oriented in a known direction in stimulus space. We call these lists *Constructed Stimulus-Vectors*, CSVs. For each CSV, stimuli were chosen so that the centroid of the first two stimuli lay near the center of the three-dimensional scaled-stimulus space. In physical units, that central value was 1.97 cyc/deg, 1.97 cyc/deg, and 1.97 radians, for vertical frequency, horizontal frequency, and phase, respectively. Additionally, the location of each stimulus in scaled stimulus space was randomly jit-

tered (by adding a Gaussian random variable with mean zero and standard deviation = 0.09) for each trial; this reduced the value of absolute phase or absolute frequency as potential cues, and discouraged subjects from forming symbolic strategies (Magnussen, Idas, & Myhre, 1998). Table 3 gives the minimum and maximum boundary values for the 3-D space from which all stimuli, study stimuli as well as probes, were drawn. Comparing the values of maximum/minimum for the various dimensions conveys the scaling factor we applied to each dimension in the stimulus space.

On some trials, all stimuli (including probes) varied along just a single dimension, e.g., horizontal frequency, with the remaining two dimensions held constant; on other trials, stimuli varied simultaneously along all three dimensions. One-dimensional and three-dimensional CSVs appeared equally often and in random order. (Figure 8 illustrates our seven different CSV orientations). In all cases, the positions of  $s_1$  and  $s_2$  were oriented in a single direction through the stimulus spacing, moving from low to high values along the various dimensions. Our reason for doing this was to limit the number of conditions so that we could obtain a sufficient number of repetitions of each list for each subject. It will be important to determine, in future studies, whether the results would differ for different directions through stimulus space.

Scaling of the stimuli, based on the application of NEMO to data from Experiment 1, predicted that the three one-dimensional CSVs would produce approximately equal performance. In creating the three-dimensional CSVs we normalized the distances along each dimension so that the Euclidean distance, scaled for each dimension in the same way as for the 1-D CSVs, would be preserved across all CSV conditions. The 3-D CSVs would produce equal performance to the 1-D CSVs if i) perceptual distances were Euclidean, ii) the space was isotropic, and iii) the covariances among the dimensions were all zero.

In addition to varying a CSV's orientation in stimulus space, we generated 12 types of lists with different similarity relations among items in a list. This arrangement is most easily seen for lists on which  $s_1$  and  $s_2$  varied along just one stimulus dimension. When that sole dimension of difference was vertical frequency,  $s_2$  was either 0.38 or 0.76 cycles/degree higher than  $s_1$ . When the study stimuli differed in horizontal frequency alone,  $s_2$  was either 0.21 or 0.42 cycles/degree higher than  $s_1$ . Finally, when phase was the sole differentiating dimension,  $s_2$  was either 0.25 or 0.50 radians greater than  $s_1$ . This arrangement, in which  $s_1$  and  $s_2$  were either relatively near to or far from one another, made it easier to determine whether episodic recognition performance was affected by the difference between study items.

For target trials we factorially manipulated the serial position of the probe (i.e., testing either  $s_1$  or  $s_2$ ), and the distance between the two study items (near or far). For the lure trials, we factorially varied the distance between the study items, and the distance between the lure and either  $s_1$  or  $s_2$ . For example, in one type of list,  $s_1$  and  $s_2$  were similar to one another, and the probe,  $p$ , was a lure more similar to  $s_2$  than to  $s_1$ ; in another type of list,  $p$  was similar to  $s_1$ , and  $s_2$  was

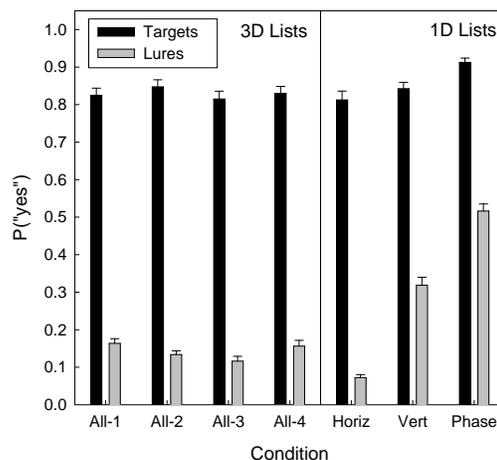


Figure 9. Hit and false alarm rates for the different CSV types in Experiment 2 (see Figure 8)

dissimilar to  $s_1$ . Altogether there were eight types of lure trials, but only 4 types of target trials. We doubled-up the number of repetitions of the target trials in order to equate the overall number of target and lure trials. It is important to bear in mind that all stimuli in a list, including any lure, were constrained to lie along some straight line in the multidimensional stimulus space.

## Results

We compared recognition memory for various CSVs. Figure 9 shows hit and false-alarm rates for CSV's of different orientations in stimulus space. Both hit and false-alarm rates are essentially the same for the four 3-D directions (leftmost bars in Figure 9). This equality suggests that, for our stimuli, there were no severe anisotropies in the psychological stimulus space.

Although recognition for 1-D vertical-differences only is equivalent to the 3-D cases, the three 1-D stimulus dimensions are not equivalently affected by interference within a list. For CSVs oriented along the dimension of horizontal frequency (whose list items did not differ in either phase or vertical frequency) both hit and false alarm rates were comparable to those of the 3-D CSVs. For CSVs oriented along the phase dimension, the false alarm rate rose by a factor of five, whereas the hit rate only rose slightly (see Figure 9). Apparently, subjects found lures that differed from the study items only in phase very tempting. A similar, but smaller effect, was observed for CSVs oriented along the horizontal frequency dimension. This may be a sign that the other, vertical, dimension drew more attention, perhaps because the harmonic and its phase were defined along that dimension. This possibility remains to be tested.

Table 3

|         | Vertical Frequency<br>(cyc/deg) | Horizontal Frequency<br>(cyc/deg) | Phase<br>(radians) |
|---------|---------------------------------|-----------------------------------|--------------------|
| Minimum | 0.704                           | 1.164                             | 1.13               |
| Maximum | 3.216                           | 2.664                             | 2.797              |

### NEMO's fit to the data from Experiment 2

Figure 10 shows that NEMO provides a good fit to the accuracy data from across the 84 CSVs. Considering that we were fitting data from so many lists with such varying performance, the observed correlation between theory and data ( $r = 0.90$ ) was respectable. Furthermore, as compared with Experiment 1, NEMO accounted for an additional 17% of the variance in subjects' performance. It is likely that part of this improved fit was a consequence of the reduced variability associated with each of the 84 CSVs (as compared with the 355 randomly generated lists of Experiment 1).

Table gives the best fitting model parameters derived from the genetic algorithm. After correcting for the differences in weightings of the three dimensions, the noise associated with storing each of the three dimensions were not significantly different from one another. Capturing the substantial effects of study-test lag on performance, NEMO reduced the contribution of the first list item by about a third in computing summed similarity.

Application of NEMO to Experiment 2 also revealed that summed similarities effect on decisions is moderated by interstimulus similarity. This is seen in the large negative value of the  $\beta$  parameter ( $-0.52$  in this Experiment) indicating that when interstimulus similarity is high subjects are less likely to say "yes" than when it is low.

As in Experiment 1, we wanted to determine whether the  $\beta$  parameter was actually responsible for a significant improvement in goodness-of-fit, we simulated a restricted version of NEMO with  $\beta = 0$ . For this probe-similarity only variant, the RMSD value increased over 40% (from 0.091 to 0.131). Utilizing the fact that models are nested, we conducted a  $\chi^2$  test for the difference in the goodness-of-fit of the two models. This comparison revealed a highly significant increase in goodness-of-fit for the generalized model ( $\Delta(df) = 1$ ,  $\Delta\chi^2 = 10.8$ ,  $p < 0.005$ ). This result, replicated in both Experiments, shows that inter-stimulus similarity is an important factor in predicting recognition memory judgments. This result suggests that we ought to rethink the classic summed similarity framework.

Table shows conditions where NEMO's deviation from the observed accuracy met the same stringent criterion as we used in Experiment 1 ( $p < 0.001$  level, Bonferroni corrected). Only 4/85 conditions met this criterion (compared with 23/355 in Experiment 1). All four of these were 'lure' CSVs in which the list items and the probe only varied along the phase dimension. This result is striking given that there were only eight CSVs of this type.

Taking a closer look at the four model failures we can see that when  $\mathbf{p}$  was near  $\mathbf{s}_1$  (in phase), NEMO was far less accurate than actual subjects, whereas when  $\mathbf{p}$  was near  $\mathbf{s}_2$ , NEMO

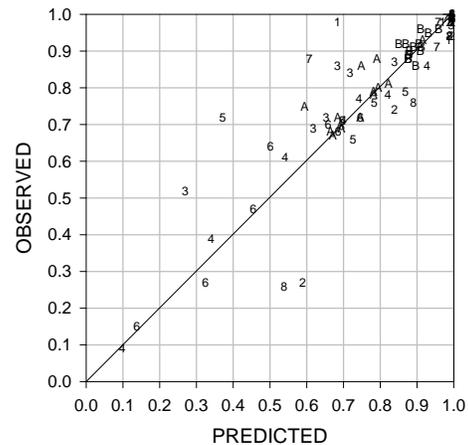


Figure 10. NEMO's attempt to fit accuracy data from Experiment 2. Observed and predicted results are plotted against one another. The symbols A and B denote targets from serial positions 1 and 2, respectively. The symbols 1–8 denote eight classes of lure trials representing the factorial combinations of (a) short and long distances between  $\mathbf{s}_1$  and  $\mathbf{s}_2$ , (b) short and long distances between  $\mathbf{p}$  and the nearest exemplar, and (c)  $\mathbf{p}$ 's that are nearer to  $\mathbf{s}_1$  vs.  $\mathbf{p}$ 's that are nearer to  $\mathbf{s}_2$ .

was far more accurate than subjects. In fact, in this last case, subjects misclassified the lure as a study item on nearly 3/4 of the trials despite the fact that the items were widely separated in stimulus space. When the lure is exactly the same distance from the well-remembered,  $\mathbf{s}_2$ , but near the poorly remembered  $\mathbf{s}_1$ , performance goes up when the model predicts that it should go down (because of increased summed similarity). Although the inter-stimulus similarity parameter,  $\beta$ , moderates the summed similarity effect to some extent, it cannot counteract this huge bias seen in subjects performance.

Table 4

| NEMO&Data | <i>NEMO's failures to fit accuracy data from Experiment 2</i> |          |             |          |          |             |        |        |           |      |
|-----------|---|----------|-------------|----------|----------|-------------|--------|--------|-----------|------|
|           | $s_1(v)$  | $s_1(h)$ | $s_1(\phi)$ | $s_2(v)$ | $s_2(h)$ | $s_2(\phi)$ | $p(v)$ | $p(h)$ | $p(\phi)$ |      |
| 0.69      | 0.98  | -0.12    | -0.12       | 1.01     | -0.12    | -0.12       | 1.80   | -0.12  | -0.12     | 0.23 |
| 0.61      | 0.88  | -0.12    | -0.12       | 1.23     | -0.12    | -0.12       | 1.63   | -0.12  | -0.12     | 0.45 |
| 0.59      | 0.27  | -0.10    | -0.10       | 1.14     | -0.10    | -0.10       | 1.92   | -0.10  | -0.10     | 2.71 |
| 0.38      | 0.72  | -0.11    | -0.11       | 1.09     | -0.11    | -0.11       | 1.88   | -0.11  | -0.11     | 0.70 |

## General Discussion

We have assessed and modeled episodic recognition memory for simple two-dimensional textures. The experimental paradigm for this work was Sternberg's (Sternberg, 1966, 1975) widely-used test of short-term recognition memory. It is important to reinforce the essential difference between episodic recognition memory and more common approaches to visual recognition. In our task, on each trial, subjects judged whether a probe item had been in the immediately preceding set of study textures. These textures varied from trial to trial. As a result of the sampling process that generated sets of study items, over time a subject would have equal exposure to all possible stimuli. Thus, subjects could not rely on a global sense that some probe was familiar or had been seen before—the typical judgment used to assess visual recognition. Instead, subjects had to judge whether the probe had been an element in the immediately preceding episode. Success in our experiments requires that subjects distinguish the most recent set of study textures from ones previously seen. In studies of non-episodic visual recognition and visual learning accumulation of experience over trials improves performance; in our task, however, memories carried over from previous episodes could undermine performance unless they were segregated from the most recent memories (Howard & Kahana, 1999, 2002).

In Experiment 1, list items were drawn randomly and without-replacement from 27 distinct positions in our 3-dimensional stimulus space. We varied list length (1–4 gratings) from trial to trial, and on trials where the probe was a list member, it was equally often chosen from among the different list positions. In Experiment 2 list length was fixed at 2 items, which, together with the probe, were chosen so that their physical features varied along a vector of known orientation in stimulus space (CSV). On some trials, all stimuli (including probes) varied along just a single dimension, with the remaining two dimensions held constant; on other trials, stimuli varied simultaneously along all three dimensions. In addition, we varied the inter-stimulus distances along the CSVs.

Experiment 1 produced a substantial list-length effect: recognition performance declined as additional items were added to the set of study items. But this result was almost entirely explained by a recency effect (superior performance for the list's most recent items). In contrast to our experiments' robust recency effects, the same recognition paradigm used with verbal stimuli tends to produce very weak recency effects (Sternberg, 1975). Moreover, with verbal stimuli the list-length effect tends to be strong even when the recency effect is absent (Sternberg, 1966). This pattern of results with

Table 5

*Best fitting parameter values for NEMO's fit to the data from Experiment 2*

| Parameter         | Value |
|-------------------|-------|
| $\sigma_{f_v}$    | 0.028 |
| $\sigma_{f_h}$    | 0.045 |
| $\sigma_{\phi_v}$ | 0.44  |
| $w_v$             | 1.0   |
| $w_h$             | 1.0   |
| $w_{\phi_v}$      | 0.066 |
| $\alpha$          | 0.65  |
| $\beta$           | -0.58 |
| $\tau$            | 10.7  |
| $c$               | 1.0   |
| RMSD              | 0.09  |

verbal stimuli can be easily understood as a consequence of rehearsal. For example, if the interval before the probe exceeds a second or so, subjects could easily use verbal or symbolic labels to cycle through some or all the study items. This could functionally rearrange the serial positions of the rehearsed items. Such a rearrangement would subvert effects associated with the serial order in which study items had been originally presented. Our visual stimuli are quite difficult to name and therefore are difficult to rehearse. This preserves the effects of recency, as seen in all of our studies, even after a delay of several seconds.

Over the years, researchers have observed many differences in memory for words and memory for pictures (e.g. Grady, McIntosh, Rajah, & Craik, 1998). Seemingly, our stimulus textures might seem more akin to pictures than to words, we hesitate to lump our results with most results on memory for pictures. Most studies of memory for pictures have used images that are sufficiently distinct from one another that subjects can give each some compact, natural verbal label. If enough time separates successive stimuli, we believe that the activation of such labels converts an image-memory task into one that is mediated by verbal labels.

## *NEMO – a Noisy Exemplar Model*

We chose stimuli constructed from simple, easily varied, visual components so that we could more readily determine the effects of interstimulus similarity on recognition memory. Considering that theories of recognition memory are almost always based on similarity computations (e.g., Hintzman, 1988; Humphreys et al., 1989; Murdock, 1997; Norman & O'Reilly, 2001; Shiffrin & Steyvers, 1997), it is sur-

prising that there are virtually no data on these effects. The reason for this is that the empirical literature is based almost exclusively on verbal materials whose similarity relations are hard to assess, and even harder to manipulate in a parametric fashion. The main purpose of our studies was to exploit the well-defined similarity relations among list items to account for recognition performance at the level of individual lists. This allows us to evaluate theories of recognition memory, which are based heavily on similarity computations among stored representations. With verbal stimuli the lack of a clear metric of interstimulus similarity makes it extremely difficult to test these theories.

Drawing on elements of both exemplar and decision bound similarity-based models, we developed NEMO, a noisy exemplar model, which we applied to data from both experiments. NEMO is a close cousin of Nosofsky's exemplar-retrieval model, GCM (Nosofsky, 1986, 1992; Nosofsky & Alfonso-Reese, 1999) and the Estes-Hintzman array models (Estes, 1986; Hintzman, 1986, 1988). Like GCM and the array models, NEMO assumes multiplexing of memories: each seen item is stored as a separate representation. Recognition decisions are then based on the summed pairwise similarity between the probe and each of the stimulus representations. As in GCM, similarity is an exponential-power function of perceptual distance (see Equation 1), and each dimension comprising the stimulus is weighted by a separate attentional factor.

NEMO departs from GCM, and other summed similarity models, in two fundamental ways. First, we assume that the stored representation of each stimulus is imperfect. To each stored representation is added a noise vector,  $\epsilon$ , whose components are mean-zero Gaussians with variances varying with the dimension and the recency of the item's occurrence. At this early stage of development of the model, we chose to make no distinction between external and internal sources of noise (Pelli & Farell, 1999). By using this noise term, we can adopt a fully deterministic decision rule (e.g., Ashby & Maddox, 1998). Variability in subjects' responses is modeled by the sampling of each item from its noisy representation. To simulate forgetting, we assume that the most recent stimulus contributes the most to the summed similarity and that earlier items contribute less and less. Stimulus-dependent noise varies across dimensions, but the rate of forgetting does not.

Second, NEMO departs from traditional summed similarity models (not just GCM, but all of the exemplar, prototype and global matching models) by assuming that subjects' responses are determined not only by the similarity between list items and the probe, but also by the similarities among list items themselves. This, together with the noisy coding of exemplars, distinguishes NEMO from most exemplar models of categorization and recognition.

We fit two versions of NEMO to data on individual lists in Experiments 1 and 2. The first, restricted version, adopted the standard assumption that the summed similarity between the probe and each of the studied items drives the recognition decision. The generalized version of NEMO assumed that inter-item similarity also contributes to recognition decisions.

In both experiments, only the generalized version of NEMO provided a good fit to data from individual lists of stimuli. The generalized model provided a significantly improved fit over the restricted model. This finding has important implications for summed similarity models of recognition (Clark & Gronlund, 1996; Estes, 1986; Hintzman, 1988; Murdock & Kahana, 1993; Nosofsky, 1992; Nosofsky & Alfonso-Reese, 1999; Shiffrin & Steyvers, 1997). In particular, summed similarity models only consider probe-to-stimulus similarities. Our findings require that the summed similarity framework be modified to include inter-stimulus similarity, as we have done in NEMO.

Our finding that interitem similarity plays a crucial role in recognition decisions involving visual textures raises the question of whether this factor plays a role in recognition memory studies more generally. In the study of verbal recognition memory the similarity relations among stimuli tend to be relatively uniform. Investigation of similarity effects in these studies generally limit themselves to comparing categorized (i.e., similar) and random (i.e., dissimilar) words lists. Furthermore, the absence, until quite recently (Howard & Kahana, 2001; Landauer & Dumais, 1997), of similarity measures for arbitrary pairs of words has made it impossible to take the analysis of recognition memory down to the level of the individual list. One could, of course, potentially generate lists of verbal stimuli with complex interstimulus similarity structures, analogous to our studies of visual textures. Designing such word lists, however, would require assumptions about the relative salience of semantic, phonological, and orthographic similarity, and there is reason to suspect that these factors differentially impact different forms of verbal episodic memory (Crowder, 1976; Murdock, 1974).

### *Limitations of NEMO*

There were several lists, especially in Experiment 1, where NEMO's predictions departed drastically from observation. These failures of the model may be due to one or more of the simplifying assumptions we have made.

First, we have assumed independence of noise along each of the stimulus dimensions (i. e., the covariances associated with  $\epsilon$  were set to zero). We compared recognition memory for CSV's of different orientations in stimulus space. Both hit and false-alarm rates were essentially the same for all four of the 3D orientations of our CSVs (leftmost bars in Figure 9). This equality suggests that there are no glaring anisotropies in the perceptual space in which the stimuli are represented. However, the overall difference between the 3-D and the 1-D cases suggests that there may in fact be interactions between the stimulus dimensions. In particular, this means that stimuli spaced at equal distances are less similar when variation is along one dimension than when it is along three dimensions. This would be consistent with a positive covariance among two or more of the three stimulus dimensions.

One compelling reason for exploring memory with stimuli like the ones we used is the good understanding afforded

of how early vision encodes such stimuli. Beginning in the late 1960's, several models have successfully linked visual discrimination directly to responses generated within ensembles of frequency and orientation-tuned, simple neurons in area V1. Olzak and Wickens (1997), among others, have pointed out the limitations of direct-access, multiple-channel models in explaining perception of real objects and scenes. The outputs of channels in early vision, they argue, must be combined in order to represent object features and segregate objects from one another. With our compound textures, such neural combination could generate higher-order configural features (e.g., crossing patterns) that could be used by vision and memory. The importance and utility of such configural cues can be assessed only from experiments that specifically manipulate such potential cues, pitting them against frequency and phase information. From a NEMO perspective, configural cues would appear as an interaction between stimulus dimensions. By setting model covariances to zero, we have implicitly assumed that such interactions are negligible.

A second, and perhaps even more serious simplifying assumption is that all three stimulus dimensions were affected by equal rates of forgetting. We made this assumption to limit the number of free parameters, but it clearly needs examination in experiments whose sets of study-test items are designed specifically for that purpose.

Third, our implementation of NEMO represents stimuli within a perceptual space whose dimensions are based on the three dimensions manipulated in our experiments, namely, vertical frequency, horizontal frequency, and relative spatial phase. The fact that we manipulated three dimensions does not guarantee, of course, that the perceptual and mnemonic representations of the stimuli were three dimensional, or, if they were, that the encoded dimensions were the ones we intended. Sensory research describes circumstances in which the number of encoded dimensions either exceeds the number of physical dimensions in the stimulus, or falls short of that number. For example, severely color deficient individuals fail to encode one or more dimensions present in chromatic stimuli, collapsing one or more dimensions of stimulus space. Or, for an opposite example, Kahana and Bennett (1994) showed that a single physical dimension, the relative phase of components in a compound grating, was processed perceptually as two dimensions, sine and cosine components, with differing sensitivities. Our data do not allow us to rule out the possibility that vision and memory actually operate on more than three dimensions, where the extra dimensions are created from non-linear combinations of the original three.

As a pre-requisite to extend NEMO to accommodate other types of visual stimuli, it will be important to verify that memory for such stimuli are as temporally robust as memory for our texture stimuli proved to be. Caution is needed because Hole (1996) showed that memory for the distance between two points diminishes within a few seconds; a similar, presumably related result occurs with memory for vernier offsets (Fahle & Harris, 1992), or for memory of the position of a previously seen single dot (Sheth & Shimojo, 2001).

Rapid loss of visual information has also been reported when subjects attempt to remember the contrast of some stimulus (Magnussen et al., 1996). Clearly, any application of NEMO to new stimulus dimensions will have to be preceded by experiments that either verify the temporal robustness of those dimensions or construct additional parameters that capture time-dependent changes in memory.

### Conclusions

We showed that NEMO, a Noisy Exemplar Model whose decisions are based on both the summed similarity of the probe to the exemplars, and the summed similarities among the exemplars, can account for recognition accuracy on individual stimulus lists. The introduction of inter-exemplar similarity into the summed similarity framework substantially improved NEMO's fit to the data from individual lists in each of two experiments. In particular, subjects' tendency to say 'yes' to a lure decreased with increasing inter-item similarity. The introduction of inter-exemplar similarity to our model represents a significant departure from the standard summed similarity framework that only considers the similarity between probes and list items. An important question for future research is to assess whether our findings regarding the role of inter-exemplar similarity can be seen for other, more complex classes of stimuli. In addition it will be important to compare NEMO with other classes of models, including those that do not rely on computations of summed similarity (e.g., Ashby & Maddox, 1998).

### References

- Ashby, F. G., & Maddox, W. T. (1998). Stimulus categorization. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (p. 251-301). Mahwah, NJ: Erlbaum.
- Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124-150.
- Blake, R., Cepeda, N. J., & Hiris, E. (1997). Memory for visual motion. *Journal Experimental Psychology: Human Perception and Performance*, *23*(2), 353-369.
- Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature*, *408*, 196-199.
- Branard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 443-446.
- Chappell, M., & Humphreys, M. (1994). An autoassociative neural network for sparse representations: Analysis and application to models of recognition and cued recall. *Psychological Review*, *101*, 103-128.
- 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.
- Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.
- De Valois, R. L., & De Valois, K. K. (1988). *Spatial vision*. New York: Oxford University Press.
- Edelman, S. (1999). *Representation and recognition in vision*. Cambridge, MA: MIT Press.

- Ennis, D. M., Palen, J., & Mullen, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology*, *32*, 449-465.
- Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*, 500-549.
- Estes, W. K. (1994). *Classification and cognition*. Oxford, U. K.: Oxford University Press.
- Fahle, M., & Harris, J. P. (1992). Visual memory for vernier offsets. *Vision Research*, *32*, 1033-1042.
- Geisler, W. S., & Albrecht, D. G. (1997). Visual cortex neurons in monkeys and cats: detection, discrimination, and identification. *Visual Neuroscience*, *14*, 897-919.
- Grady, C. L., McIntosh, A. R., Rajah, M. N., & Craik, F. I. M. (1998). Neural correlates of the episodic encoding of pictures and words. *Proc. Nat. Acad. Sci., USA*, *95*, 2703-2708.
- Graham, N. V. S. (1989). *Visual pattern analyzers*. New York: Oxford University Press.
- Hintzman, D. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, *93*, 411-428.
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in multiple-trace memory model. *Psychological Review*, *95*, 528-551.
- Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review*, *94*, 341-358.
- Hole, G. J. (1996). Decay and interference effects in visuospatial short-term memory. *Perception*, *25*, 53-64.
- Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923-941.
- Howard, M. W., & Kahana, M. J. (2001). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*, 85-98.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*.
- 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., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology*, *33*, 36-67.
- Kahana, M. J., & Bennett, P. J. (1994). Classification and perceived similarity of compound gratings that differ in relative spatial phase. *Perception & Psychophysics*, *55*(6), 642-656.
- Kahana, M. J., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (p. 322-384). Cambridge, MA: MIT Press.
- Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., Ganis, G., Sukel, K. E., & Alpert, N. M. (1999). The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science*, *284*, 167-170.
- Landauer, T. K., & Dumais, S. T. (1997). Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.
- Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*, 89-94.
- Maddox, W. T., & Ashby, F. G. (1996). Perceptual separability, decisional separability, and the identification-speeded classification relationship. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 795-817.
- Magnussen, S. (2000). Low-level memory processes in vision. *Trends in Neuroscience*, *23*, 247-251.
- Magnussen, S., & Greenlee, M. (1999). The psychophysics of perceptual memory. *Psychological Research*, *62*, 81-92.
- Magnussen, S., Greenlee, M. W., Asplund, R., & Dyrnes, S. (1990). Perfect visual short-term memory for periodic patterns. *European Journal of Cognitive Psychology*, *2*, 345-362.
- Magnussen, S., Greenlee, M. W., Asplund, R., & Dyrnes, S. (1991). Stimulus-specific mechanisms of visual short-term memory. *Vision Research*, *31*, 1213-1219.
- Magnussen, S., Greenlee, M. W., & Thomas, J. P. (1996). Parallel processing in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 202-212.
- Magnussen, S., Idas, E., & Myhre, S. H. (1998). Representation of orientation and spatial frequency in perception and memory: a choice reaction-time analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 707-718.
- Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 3-23.
- McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 294-317.
- Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.
- Murdock, B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Erlbaum.
- Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*, 839-862.
- Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list strength effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 689-697.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308-313.
- Norman, K. A., & O'Reilly, R. C. (2001). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *submitted*.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.
- Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 363-394). Hillsdale, New Jersey: Erlbaum.
- Nosofsky, R. M., & Alfonso-Reese, L. A. (1999). Effects of similarity and practice on speeded classification response times and accuracies: Further tests of an exemplar-retrieval model. *Memory & Cognition*, *27*, 78-93.
- Nosofsky, R. M., & Palmeri, T. J. (1998). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.
- Olzak, L. A., & Thomas, J. P. (1999). Neural recoding in human pattern vision: Model and mechanisms. *Vision Research*, *39*, 231-256.

- Olzak, L. A., & Wickens, T. (1997). Discrimination of complex patterns: Orientation information is integrated across spatial scale; spatial-frequency and contrast information are not. *Perception*, *26*, 1101-1120.
- Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.
- Pelli, D. G., & Farell, B. J. (1999). Why use noise. *Journal of the Optical Society of America*, *16*, 647-53.
- Pelli, D. G., Robson, J. G., & Wilkins, A. J. (1988). Designing a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences*, *2*, 187-199.
- Sheth, B. R., & Shimojo, S. (2001). Compression of space in visual memory. *Vision Research*, *41*, 329-341.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM — retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*, 145.
- Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652-654.
- Sternberg, S. (1975). Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology*, *27*, 1-32.
- Teller, D., & Palmer, J. (1998). *Psychophysics, mit encyclopedia of cognitive science*. Cambridge, Mass: MIT Press.
- Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*, 68-111.
- Wilson, H. R., & Wilkinson, F. (1997). Evolving concepts of spatial channels in vision: from independence to nonlinear interactions. *Perception*, *26*(8), 939-960.