Memory & Cognition
2006, 34 (1), 112-124

Exemplar similarity, study list homogeneity, and short-term perceptual recognition

ROBERT M. NOSOFSKY and JUSTIN KANTNER
Indiana University, Bloomington, Indiana

Kahana and Sekuler (2002) conducted short-term perceptual recognition experiments and modeled the data with a noisy exemplar similarity model. They found model-based evidence that list homogeneity (i.e., the degree to which exemplars on a study list are similar to one another) exerted a significant impact on recognition performance—a finding that is not predicted by standard global familiarity models. A potential limitation of their experiments is that they tested complex stimuli in which psychological similarities among exemplars may have been misspecified. Also, the relative importance of list homogeneity was not compared with that of alternative forms of parametric variation in the model. We conducted conceptual replications of their experiments, using a simpler set of stimuli in which interexemplar similarities could be more precisely measured. Extensive model-based comparisons reveal, in accord with the results of Kahana and Sekuler, strong evidence for a role of list homogeneity on *old–new* recognition performance. We suggest that subjects systematically adjust their response criteria on the basis of the homogeneity of the study list items.

According to exemplar models of perceptual recognition memory, observers make *old–new* recognition judgments by assessing the overall familiarity of test items (Gillund & Shiffrin, 1984; Hintzman, 1988; Nosofsky, 1988, 1991). Items on a study list are represented as individual exemplars in memory. Presentation of a test item leads to the activation of the stored exemplars. This activation is determined jointly by the similarity of the test item to the stored exemplars and by the strength with which the individual exemplars are stored in memory. The greater the overall activation, the more familiar is the test item, and so the greater is the probability with which the observer will judge the item to be old.

Exemplar models have had a long history of success in predicting quantitative details of perceptual recognition performance. Importantly, such models have been capable of predicting fine-grained differences in *old–new* recognition probabilities as a function of fine-grained differences in similarities among items. Thus, such models have been reasonably successful at predicting not only overall performance differences across conditions, but also the probability with which individual items will be judged as old or new (e.g., Busey & Tunnicliff, 1999; Lamberts, Brockdorff, & Heit, 2003; Nosofsky, 1991; Shin & Nosofsky, 1992; Zaki & Nosofsky, 2001).

This work was supported by National Institute of Mental Health Grant R01 MH48494. The authors thank Michael Kahana for extensive discussions and suggestions related to the work reported in this article, Robert Sekuler for providing the stimuli used in the similarity-scaling study and for his encouragement of the research, and three anonymous reviewers for their criticisms of an earlier version of the article. Correspondence concerning this article should be addressed to R. M. Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405 (e-mail: nosofsky@indiana.edu).

In recent work, however, Kahana and Sekuler (2002) observed an important limitation of exemplar models of perceptual recognition. Specifically, they observed conditions in which the sole use of a summed-activation rule failed to account adequately for the details of individual-item recognition performance. Instead, Kahana and Sekuler obtained evidence that, beyond the similarity of test items to the list exemplars, the similarity of list exemplars to one another also exerted an important impact. This finding is of great potential significance, because it points to a factor influencing recognition judgments that was previously unknown and was not incorporated into extant models. The primary purpose of the present research was to pursue this important finding of Kahana and Sekuler and to investigate it in greater detail.

Review of Kahana and Sekuler's (2002) Experiments

Kahana and Sekuler (2002) employed a variant of the classic Sternberg (1966) short-term memory-scanning paradigm in their research. On each trial, a short list of study items was presented. Following this list, a test probe was presented, and the subjects were required to judge whether the probe was old (presented on the study list) or new (not presented). Whereas Sternberg and others used discrete, nonconfusable stimuli (e.g., alphanumeric characters) as study items, Kahana and Sekuler used confusable stimuli embedded in a continuous multidimensional similarity space. Given the confusable nature of the stimuli, the focus of their research was on predicting recognition choice probabilities. (By contrast, in Sternberg's classic work, accuracy was near ceiling, and the focus was on predicting how response times varied as a function of experimental conditions.)

Kahana and Sekuler (2002) used a *noisy exemplar model* (NEMO) for modeling their data. NEMO combines assumptions from Nosofsky's (1986, 1991) *generalized context model* (GCM) of classification and recognition with perceptual/memory noise ideas that form part of decision boundary theory (e.g., Ashby & Townsend, 1986; Ennis, 1992). In brief, according to NEMO, study items are represented as noisy exemplars in memory. Test probes give rise to an overall familiarity measure based on their similarity to the noisy exemplars. If the familiarity exceeds a criterion, the observer judges the item to be old; otherwise, the observer judges the item to be new.

We will start by describing a baseline version of the formal model. According to NEMO, across trials, each exemplar gives rise to a multivariate normal distribution of points in a multidimensional similarity space (e.g., Ashby & Townsend, 1986; Ennis, 1992; Nosofsky, 1997). This distribution constitutes the noisy memory representation associated with exemplar j. Exemplar distribution j has mean μ_{jm} on dimension m and has memory variance σ_m^2. For simplicity, it is assumed that there are no interdimensional correlations. On a given trial, the memory of a study exemplar is represented by a single point, x_{jm}, which is selected randomly from this multivariate distribution. Suppose that test probe i is presented. The probe is represented by a single point with location $x_{im} = \mu_{im}$. (For simplicity, it is assumed here that sensory noise is negligible, so sensory variances are not estimated. The contribution of sensory noise is absorbed by the memory variance estimates.)

The similarities between test probes and the noisy exemplar representations are computed in the same manner as in the GCM. Specifically, the distance between probe i and exemplar representation j is given by a weighted Euclidean distance metric,

$$d_{ij} = \left[\sum w_m \cdot \left| x_{im} - x_{jm} \right|^2 \right]^{1/2}, \tag{1}$$

where w_m ($0 \leq w_m$) is the attention weight given to dimension m. The similarity between probe i and exemplar j is then given by

$$s_{ij} = \exp\left(-c \cdot d_{ij}^{\alpha}\right), \tag{2}$$

where c is an overall sensitivity parameter that determines the rate at which similarity decreases with distance and α determines the shape of the similarity gradient.

According to NEMO, the observer sums the similarity of the probe to all of the study list exemplars (cf. Nosofsky, 1988, 1991),

$$S = \sum_{p=1}^{P} M_p \cdot s_{ij}(p). \tag{3}$$

In this equation, $s_{ij}(p)$ denotes the similarity of probe i to exemplar j, where exemplar j resides in serial position p of the study list and where P is the number of exemplars in the list. M_p denotes the memory strength associated with the exemplar that resides in serial position p. In general, the more recently a study exemplar has been presented, the greater is its memory strength. Finally, if the summed similarity S exceeds a criterion k, $S > k$, the observer responds that the probe is *old*; otherwise, the observer responds *new*. Note that, according to NEMO, *old–new* recognition decisions will be probabilistic across trials, even if the same study list and probe are presented. The reason is that the study exemplars give rise to noisy representations in the multidimensional similarity space, so the summed-similarity term S varies probabilistically.

In their experiments, Kahana and Sekuler (2002) found that this baseline version of NEMO yielded fair quantitative predictions of the *old–new* recognition results, but there were important shortcomings in the quality of the fit as well. An improved fit to the data was achieved by augmenting the standard summed-similarity rule of exemplar models with a *list homogeneity* parameter. Specifically, the homogeneity (H) of a given study list was defined by computing the average similarity of each study exemplar to every other study exemplar:

$$H = \sum_{i=1}^{P-1} \sum_{j=i+1}^{P} s_{ij}(p) / [P \cdot (P-1)/2]. \tag{4}$$

According to the modified decision rule in the NEMO model, an observer judges a probe to be old if

$$S + \beta \cdot H > k, \tag{5}$$

where β is a freely estimated parameter. In their experiments, Kahana and Sekuler found that this additional free parameter yielded significantly improved fits to their *old–new* recognition data. Furthermore, the best-fitting β parameter was negative in value. The latter result implies that, all other things being equal, subjects respond *old* less often when study lists are highly homogeneous, rather than heterogeneous.

Motivation for the Present Research

Although the evidence for the role of list homogeneity on recognition performance is intriguing, we felt that several aspects of this evidence needed to be pursued. Our first concern involved the type of stimuli that Kahana and Sekuler (2002) used in their experiments. The stimuli were compound sinusoidal grating patterns. The stimuli varied orthogonally along three physical dimensions: frequency of vertical grating (three levels), frequency of horizontal grating (three levels), and relative phase of the components' spatial positions (three levels), for a total of 27 stimuli. In their formal modeling analyses, Kahana and Sekuler assumed that the positions of the stimuli in psychological space matched this physical specification. Although this assumption provides a reasonable starting point, it is well known that psychological codings of stimuli do not always match the physical specifications provided by the experimenters (cf. Lockhead, 1972; Shepard, 1962).

Indeed, we have verified with psychological-scaling procedures that there are important emergent dimensions that influence subjects' similarity judgments of these compound gratings stimuli (see Appendix A). For example, one emergent dimension involves the extent to

which the frequency of the horizontal and vertical components match, thereby creating a square-shaped versus non–square-shaped texture grid. Thus, suppose that one stimulus has low-frequency gratings on both its horizontal and vertical components, whereas a second stimulus has high-frequency gratings on both components. Although the stimuli are separated by a large distance in the physically specified space, they may nevertheless be judged as similar because both have a square-shaped texture grid.

Our view is that, to the extent that the psychological similarity structure of the set of stimuli is misspecified, it potentially brings into question Kahana and Sekuler's (2002) evidence for the role of the homogeneity parameter on recognition judgments. For example, the homogeneity parameter could be playing the role of a "patch" for correcting misspecified summed similarities.

To document this concern, we conducted various simulations of performance in Kahana and Sekuler's (2002) paradigm and fitted NEMO to these simulated data. Although a detailed presentation would go beyond the scope of this article, the nature of the investigations was as follows. First, we used the baseline version of NEMO (i.e., the version with $\beta = 0$) to simulate performance. In these baseline model simulations, similarities among exemplars were computed by using a four-dimensional scaling solution for the gratings stimuli that we derived in our psychological-scaling study (Appendix A). Next, we fitted NEMO to these simulated data. However, in conducting these fits, rather than using the "true" similarities computed from the scaling solution, we fitted the model by using the three-dimensional physical description of the stimuli that was assumed in Kahana and Sekuler's analyses. Furthermore, we allowed the β parameter to vary freely in conducting the fits. Interestingly, despite the fact that β had been held fixed at zero in generating the simulated data, the model yielded significantly better fits when β was allowed to vary freely. Furthermore, the estimated value of β was strongly negative, just as Kahana and Sekuler had observed. Thus, the possibility that β might be capturing unexplained response variance resulting from misspecified similarities is not an idle concern.

Accordingly, the first major purpose of the present research was to conduct a conceptual replication of Kahana and Sekuler's (2002) experiments, except that we used an alternative set of stimuli for which psychological similarity relations are well understood. In particular, we conducted a conceptual replication of their experiments, using a set of Munsell colors as stimuli. Extensive psychological-scaling work indicates that similarity relations among the color stimuli are extremely well described in terms of differences along the dimensions of brightness, saturation, and hue. The key question was whether or not we would find evidence similar to that in Kahana and Sekuler for a role of list homogeneity in observers' recognition judgments. We were also interested in testing the overall quantitative accuracy of the model in this domain in which interexemplar similarity could be precisely measured.

A related purpose of this research was to investigate whether or not other parameters would vary systematically as a function of study list conditions. Importantly, list homogeneity might be only one factor that plays a significant role. For example, it seems plausible that the variance of the exemplar-based memory representations might increase with increased study lag (i.e., subjects would have more sensitive memories for recently presented exemplars than for ones presented earlier on the study list).

Because the present paradigm involves the collection of a massive amount of data, it seems likely that numerous forms of parametric variation would lead to statistically significant improvements in fit. However, if the effect of allowing the list homogeneity parameter to vary were large relative to other forms of parametric variation, it would point to the importance of this factor in influencing *old–new* recognition judgments.

EXPERIMENT 1

The stimuli used in Experiment 1 were a set of computer-generated colors, derived from the Munsell collection, that varied along three dimensions: brightness (three levels), saturation (three levels), and hue (three levels). The dimension values were combined orthogonally to create a set of 27 stimuli. On each trial, a subject was presented with a study list of between one and four randomly chosen colors. A probe color was then presented, and the subject judged whether it was old (i.e., a member of the current study list) or new. There was an equal number of one-item, two-item, three-item, and four-item study lists. For each list type, half the probe items were old, and half were new. Responding was unspeeded, and the subject was instructed to be as accurate as possible.

Method

Subjects. The subjects were 94 undergraduates at Indiana University, who participated in partial fulfillment of an introductory psychology course requirement. All claimed to have normal color vision. Small monetary bonuses were offered for good performance, to help ensure subject motivation.

Stimuli. The stimuli were 27 computer-generated colors derived from the Munsell collection. The dimensional structure of the colors was analogous to that of the physically manipulated dimensions of the gratings used in Kahana and Sekuler's (2002) experiments. The original Munsell colors varied along three dimensions: hue (values: 7.5 purple-blue, 2.5 purple-blue, and 7.5 blue), saturation (values: 6, 8, and 10), and brightness (values: 4, 5, and 6). The dimension values were combined orthogonally to yield the $3 \times 3 \times 3$ stimulus set. The Munsell colors were scanned into the computer for purposes of stimulus presentation. Previous scaling work conducted by Zaki and Nosofsky (2001) indicated that the scanned colors match well the dimensional structure of the original Munsell colors. (Although the computer-generated colors are unlikely to form a perfect $3 \times 3 \times 3$ grid in the underlying psychological space, they do not give rise to the same types of emergent dimensions as those for the gratings.) The red–green–blue (RGB) values for the 27 scanned colors are reported in Appendix B. Each color occupied a 2×2 in. square in the center of a computer screen, displayed against a white background.

There was a total of 360 lists, with 90 lists each of lengths 1–4. Each list was followed by a probe item. Within each list length, half of the probe items were old, and half were new. The one-item, two-item, three-item, and four-item study lists were created by sampling randomly from the pool of 27 colors, with the restriction that no color appeared in the same list twice. Old probes matched a ran-

domly selected list item. New probes were selected randomly from the pool of colors not included in the list. The same 360 lists were used for all the subjects.

Procedure. The 360 lists were presented in a different randomized order to each subject. To enable the subjects to keep track of their progress, each trial was preceded by the trial number, displayed in the center of the screen for 1 sec. The screen was then blank for 1 sec, after which list presentation began. Each list item was presented for 1 sec, with a blank 1-sec interstimulus interval separating the items. Following the final list item, a focal point ("x") appeared in the center of the screen for 1 sec, and then the probe appeared with the question, "Was this color on the preceding list? (yes/no)." The subjects received immediate feedback ("Correct!" or "Incorrect!") following each response. Two unscored practice lists were presented at the start of the experiment, and breaks followed every 90 trials.

Results

Prior to modeling the data, we created a histogram of the overall percentage of correct responses achieved by each of the subjects. On the basis of an inspection of the histogram, we deleted the data of 14 subjects who had performed very poorly (less than 61% correct overall). Our subsequent analyses are based on the data from the 80 remaining subjects.

Although the primary aim was to model quantitatively the recognition probabilities associated with the individual lists, we will start by reporting some general characteristics of the data. The mean probability of correct recognition decisions is shown as a function of list length, lag, and probe status (old vs. new) in Figure 1. Lag is defined as the number of items that intervene between old test probes and their occurrence on the study list. As can be seen in the figure, for all list lengths, recognition probabilities increased with decreasing lag. For old probes, there was little if any effect of list length once lag was taken into account. However, there was a clear effect of list length on accuracy

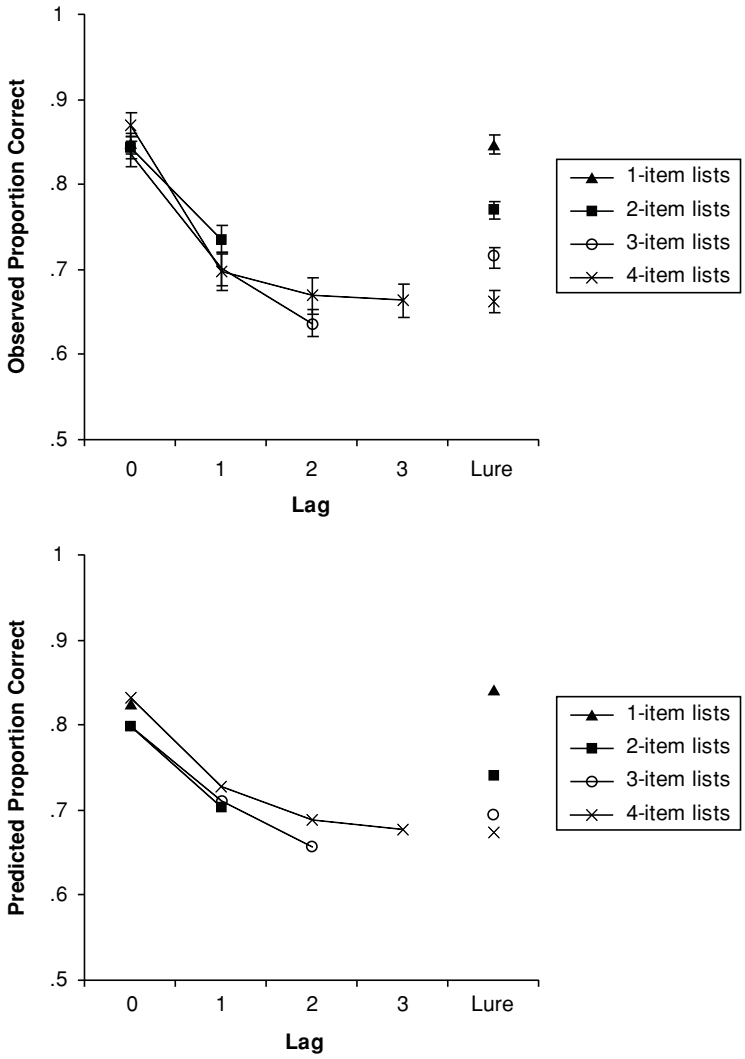

Figure 1. Experiment 1: mean proportion correct as a function of lag, list length, and probe status (old or new). Top panel, observed data; bottom panel, predictions from NEMO.

for lures, with greater correct rejections being associated with shorter list lengths. This pattern of results replicates the pattern that Kahana and Sekuler (2002) observed with their compound gratings stimuli. Furthermore, as will be seen, this pattern of results is as predicted by NEMO.

Formal Modeling Analyses

Our central goal was to use NEMO to predict the mean probability with which the subjects made *old* judgments for each of the individual 360 test lists. We fitted NEMO to these data by conducting extensive computer searches for the best-fitting free parameters. For any given set of parameters, 1,000 simulations were conducted for each test list, to generate the predicted probabilities from the model. The computer algorithm searched for the values of the free parameters that provided a maximum-likelihood fit to the data.

We fitted NEMO to the recognition data by using the classic Munsell scaling solution associated with the colors. According to the Munsell scaling, the colors are evenly spaced along each of their component dimensions. We coded the values along each dimension with the magnitudes 1, 2, and 3. These magnitudes correspond to the means of the memory distributions associated with each exemplar—that is, the values of μ_{jm}. The relative discriminability of each of the dimensions is modeled in terms of the estimated values of the attention weight and dimensional variance parameters in NEMO.

We fitted different versions of NEMO to the data by either constraining or allowing to vary freely key parameters in the model. By comparing systematically the quantitative fits yielded by the different versions of NEMO, we can gain information regarding the importance of the various free parameters. Because of the massive amount of data collected, allowing any parameter to vary freely yielded statistically significant improvements in fit when standard likelihood-ratio testing methods were used. Therefore, we will dispense with reporting these types of statistical comparisons and will focus instead on a description of which free parameters seem to do the lion's share of the work in accounting for the data. The results from the different versions of NEMO, including the best-fitting parameters and model summary fits, are reported in Table 1.

The *core* version of the model (Version A) includes the following free parameters: the overall sensitivity parameter c (Equation 2), the attention weights w_m (Equation 1), the dimensional memory variances σ_m^2, the exemplar-memory strengths associated with each lag (M_L), an overall response criterion parameter k, and the list homogeneity parameter β. In this core version of the model, we hold fixed the α similarity gradient parameter (Equation 2) at $\alpha = 1$, which yields the common assumption of an exponential relation between similarity and psychological distance (Shepard, 1987). Note in addition that the core version assumes a single response criterion parameter (k) for all list lengths, as well as constant memory variance regardless of lag. We will consider the results of relaxing these assumptions below.

A scatterplot of the observed against the predicted recognition probabilities for each of the individual 360 lists is shown in Figure 2. The predictions of the main trends are shown along with the observed data in Figure 1. This core version of NEMO provides an excellent quantitative fit to the recognition data, accounting for 92% of the response variance associated with the individual study lists. The fits here are substantially better than the ones achieved by Kahana and Sekuler (2002; 64% of the individual-list response variance accounted for). One likely contributing

Table 1
Experiment 1: Best-Fitting Parameters and Summary Fits From the Different Versions of NEMO

	Model Version					
Parameters	A	B	C	D	E	F
c	5.571	5.074	5.649	5.461	5.651	5.861
w_2	0.065	0.087	0.150	0.136	0.150	0.150
w_3	0.293	0.293	0.390	0.556	0.430	0.390
σ_1	0.379	0.441	0.336	0.327	0.325	0.324
σ_2	1.045	0.807	0.985	1.098	1.003	0.973
σ_3	0.692	0.642	0.796	0.771	0.796	0.821
M_2	1.304	1.000	2.051	1.620	1.830	1.210
M_1	1.685	1.000	3.035	3.046	2.763	1.668
M_0	3.119	1.000	7.307	18.923	6.378	3.432
k	0.044	0.025	0.054	0.267		0.027
β	−3.611	−2.733	0.000	0.000	0.000	0.000
Fits						
$-\ln L$	1,798.2	2,058.2	1,969.2	1,879.5	1,919.7	1,939.9
SSD	2.797	4.193	3.737	3.163	3.497	3.438
% Var.	92.0	88.0	89.3	90.9	90.0	90.1

Note—Because only the relative values of the attention weights and memory strengths can be measured, the values of w_1 and M_3 were held fixed at 1 in all fits. In Version D, $\alpha = 3.122$. In Version E, $k_1 = .032$, $k_2 = .044$, $k_3 = .046$, and $k_4 = .048$. In Version F, $v_0 = .925$, $v_1 = .991$, $v_2 = 1.022$, and $v_3 = 1.128$.

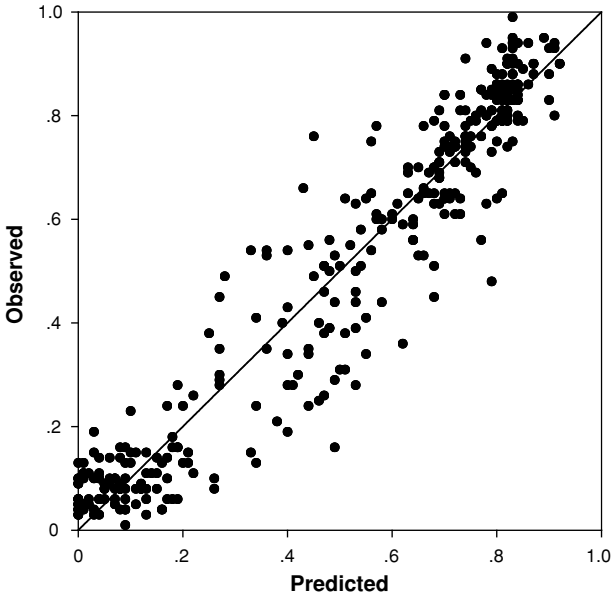

Figure 2. Experiment 1: observed against predicted probabilities of *old* recognition responses for each of the 360 lists.

reason is that interexemplar similarities are more precisely measured with the current set of stimuli. Another related possibility is that there may have been more intersubject variability in perception of the gratings stimuli and that this increased variability led to noisier data in Kahana and Sekuler's study.

The lag-dependent memory strength parameters, of course, play a major role in achieving the good fits. This mechanism allows NEMO to account for the large effect of memory lag that is seen in Figure 1. As can be seen in Table 1, the estimated memory strengths increase systematically with decreasing lag. A special case of the model (Version B), in which all memory strength parameters are held fixed at 1.0, provides a dramatically worse fit to the recognition data (see Table 1).

Note that in addition to predicting the lag functions, NEMO also captures well the overall effect of list length on the accuracy rates (see Figure 1). False alarms tend to increase as a function of list length for two reasons. First, all other things being equal, the greater the number of items on the study list, the greater will be the summed similarity of the test probe to the memory set items. Second, with increasing list length, there is an increased probability that at least one memory set exemplar will be highly similar to the test probe. The effect of list length is smaller for old test probes. The reason is that the summed similarity is dominated by the match between the test probe and its own memory representation from the study list.

The main question of interest concerns the role of the list homogeneity parameter β. First, note from Table 1 that the best-fitting value of the list homogeneity parameter was $\beta = -3.661$. Thus, as was found by Kahana and Sekuler (2002), the results suggest that, all other things being equal, observers are more willing to accept test probes as old when they experience heterogeneous, rather than homogeneous, study lists. To bring out the importance of the list homogeneity parameter, we fitted a special case of NEMO (Version C) in which the β parameter was held fixed at zero. The quantitative fit of this special case model is substantially worse than that of the core version of the model: The (negative) log-likelihood statistic increases from 1,798.2 to 1,969.2, and the sum-of-squared deviations (*SSD*) between predicted and observed recognition probabilities increases from 2.797 to 3.737.

To bring out further the importance of this list homogeneity parameter, we will report the fits of other versions of NEMO as sources of comparison. In Version D, we hold β fixed at zero but allow the α similarity gradient parameter to vary freely. As can be seen in Table 1, making allowance for a more flexible similarity gradient does not improve the fit nearly as much as does taking into account the role of list homogeneity.

In Version E, we hold β fixed at zero but make allowance for different response criterion settings as a function of list length. Note that adding the three additional response criterion parameters does not improve the fit as much as does adding the single list homogeneity parameter. (Note as well that this version of NEMO generalizes the version tested by Kahana and Sekuler [2002], which made the strong assumption that subjects adopted ideal-observer criteria for each individual list length.) As yet another source of comparison, we fitted a version of NEMO (Version F) in which the memory variances were allowed to depend on lag. Specifically, the memory variance along dimension m at lag L, $\sigma_m^2(L)$, was assumed to be given by

$$\sigma_m^2(L) = v_L \cdot \sigma_m^2, \tag{6}$$

where v_L is a lag-related variance multiplier. (As is the case in the multiple response criterion version, three additional free parameters are incorporated here.) Again, however, making allowance for changes in memory variance as a function of lag does not improve the quantitative fit as much as does including the single β list homogeneity parameter (compare the quantitative fits of Versions A and F).

Finally, we also fitted elaborated models in which β was allowed to be a free parameter and in which the similarity gradient (α), variance multiplier, and list-length–specific criterion parameters were allowed to vary freely as well. Although adding these free parameters led to statistically significant improvements in fit, in no case was the fit very much better than what was achieved by the core version of the model.

Discussion

In sum, all of these quantitative comparisons among the different versions of NEMO point to the relative importance of the list homogeneity parameter in achieving good fits to the *old–new* recognition data. The experimental and modeling results provide an important conceptual replication and further documentation of Kahana and Sekuler's (2002) findings. First, by documenting a role of list homogeneity in the present stimulus domain, we remove the concern that Kahana and Sekuler's results may have

involved an artifact due to misspecifying underlying psychological similarity relations. Second, by including comparison fits of a wide variety of alternative models, our results point to the relative importance of list homogeneity in influencing subjects' *old–new* recognition judgments.

EXPERIMENT 2

In Experiment 1, we found evidence for a role of list homogeneity in terms of an overall improved quantitative fit to the data. However, the study lists were chosen randomly, and there were no focused comparisons to help bring out the role of the list homogeneity parameter. The purpose of Experiment 2 was to include study lists in which homogeneity was explicitly manipulated so as to achieve such comparisons.

In Experiment 2, we used only two-item study lists. Again, there was a total of 360 lists that were tested. Half of the lists were random lists that were generated by using the same methods as those in Experiment 1. The other half of the lists were *critical* lists in which homogeneity was explicitly manipulated. The structure of the critical lists is illustrated schematically in Figure 3. For *high-homogeneity* lists, the two study items were always of the same hue and were immediately adjacent in the brightness–saturation plane. For high-homogeneity lists in which the test probe was new, the probe was immediately adjacent to one study item and diagonally adjacent to the other study item. For high-homogeneity lists in which the test probe was old, the probe matched one of the study items. For *low-homogeneity* lists, the two study items were always of different hues and were far apart on the dimensions of brightness and saturation as well. For low-homogeneity lists in which the test probe was new, the probe was immediately adjacent to one of the study items; for low-homogeneity lists in which the test probe was old, the probe matched one of these study items.

If list homogeneity influences performance in the hypothesized manner, the results from the critical lists should provide dramatic evidence of the effect. Note that, all other things being equal, the summed similarity of the test probes to the study exemplars is much greater for the high-homogeneity lists than for the low-homogeneity lists. Thus, without including the list homogeneity parameter, the summed-similarity exemplar model predicts much higher hit and false alarm rates for high-homogeneity lists than for low-homogeneity lists. The differences are predicted to be much smaller and could possibly even reverse direction, according to the core version of NEMO.

Method

Subjects. The subjects were 60 Indiana University undergraduates who participated in partial fulfillment of an introductory psychology course requirement. All reported having normal color vision. Monetary bonuses were again offered to motivate high accuracy.

Stimuli. The same 27 colors as those in Experiment 1 were used. There were 360 two-item lists constructed from these colors according to three different list types: high homogeneity, low homogeneity, and random. High-homogeneity lists contained two study items (S1 and S2) from the same hue region, equivalent in either saturation or brightness value and immediately adjacent on the nonequivalent dimension (see Figure 3). Probe items for high-homogeneity lists came from the same hue region as the list items. Old probes matched one of the list items; new probes were immediately adjacent to one of the list items in either brightness or saturation value and were diagonally adjacent to the other list item.

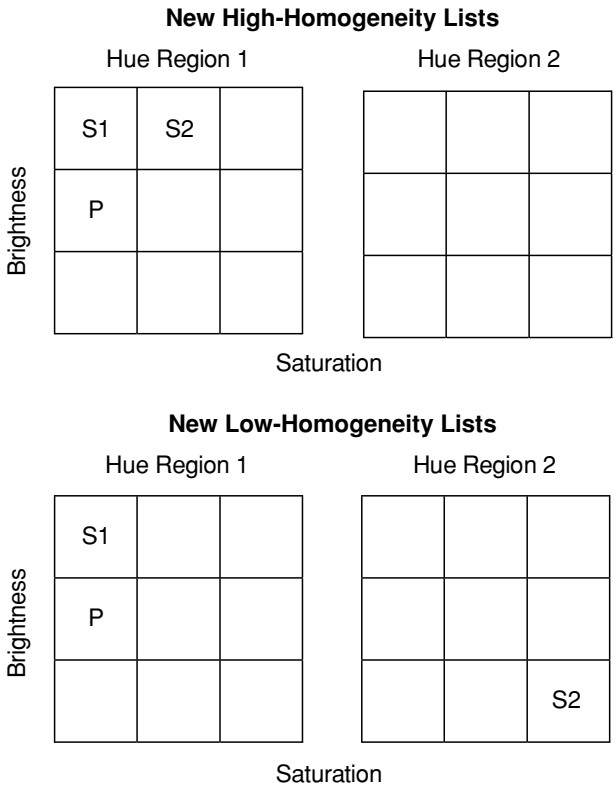

Figure 3. Experiment 2: schematic illustration of the design. S1, Study Item 1; S2, Study Item 2; P, probe item. Hue Region 3 is not shown (no colors would appear in the third hue region for either type of list). For lists with *old* test probes, the probe would be identical to the S1 illustrated in the figure.

Low-homogeneity lists were yoked to high-homogeneity lists (see Figure 3). Each low-homogeneity list was created by deleting one of the items (S2) from an existing high-homogeneity list and replacing it with an item from a separate hue region and as far away from the other list item (S1) as possible in brightness and saturation value. Old probes matched S1, whereas new probes were immediately adjacent to S1.

For both high- and low-homogeneity lists, S1 came from each brightness and saturation coordinate equally often across hue regions. For each list, the positions of S2 and the probe (relative to S1) were chosen at random, within the constraints of the design above. Serial position of S1 and S2 on the lists was also chosen randomly. Note that this design ensures that, across trials, the probe can occupy any one of the 27 locations in the color space with equal probability. Thus, the subjects would not be able to learn a strategy of always responding in some manner on the basis of the absolute location of the probe.

There was a total of 90 high-homogeneity and 90 low-homogeneity lists. A total of 180 random two-item lists—generated by the method used in Experiment 1—was also constructed. All the subjects were tested on the same lists.

Procedure. The trial structure and procedure were identical to those in Experiment 1.

Results

On the basis of an inspection of the overall percentage of correct responses, we deleted the data of 6 subjects who had performed poorly (less than 55% correct responses overall). Our subsequent analyses are based on the data from the 54 remaining subjects.

The mean probability with which the subjects endorsed test items as old is reported as a function of list type in Table 2. We analyzed the results from the critical lists by using a 2 × 2 ANOVA with homogeneity (high vs. low) and probe type (old vs. new) as factors. The subjects endorsed old probes with higher probability ($M = .729$) than they endorsed new probes ($M = .417$) [$F(1,53) = 308.0$, $MS_e = 0.017$, $p < .001$]. The subjects also endorsed probes from high-homogeneity lists with higher probability ($M = .610$) than they endorsed probes from low-homogeneity lists ($M = .537$) [$F(1,53) = 44.9$, $MS_e = 0.006$, $p < .001$]. The interaction between probe type and list homogeneity was also significant [$F(1,53) = 16.9$, $MS_e = 0.005$, $p < .001$], reflecting that the boost in hit rates exceeded the boost in false alarm rates.

The general pattern of results described above is in accord with the predictions of standard summed-similarity exemplar models of recognition. The most important observation, however, is that the difference in recognition probabilities associated with low- versus high-homogeneity lists appears to be rather small, especially for new probes. This result provides an immediate clue as to the importance of the list homogeneity parameter. Without this parameter, standard summed-similarity models should predict far greater false alarm rates for the high-homogeneity lists than for the low-homogeneity ones: Probes are highly similar to two study exemplars in the case of the high-homogeneity lists but are highly similar to only a single study exemplar in the case of the low-homogeneity lists. We will turn now to the formal modeling analyses to check this intuition.

Theoretical Analyses

We fitted the different versions of NEMO to the *old–new* recognition data in the same manner as that described in Experiment 1. The results of the formal modeling analyses (best-fitting parameters and summary fits) are reported in Table 3. The key comparison of interest is between the core version of NEMO (Version A) and the version in which the β parameter is held fixed at zero (Version C). As can be seen, the fit of the restricted version in which $\beta = 0$ is dramatically worse than that of the core model. The (negative) log-likelihood statistic increases from 1,399.8 to 1,631.0, and the *SSD* between predicted and observed recognition probabilities increases from 4.089 to 6.119. Furthermore, comparison with the other versions of NEMO indicates that the β parameter again plays a substantially more important role than do alternative parametric variations of the model.

Table 2
Experiment 2: Mean Observed and Predicted Recognition Probabilities for the Different Types of Test Lists

List Type	Observed	NEMO (β free)	NEMO ($\beta = 0$)
Old, high homogeneity	.785	.758	.799
Old, low homogeneity	.673	.670	.611
New, high homogeneity	.434	.465	.515
New, low homogeneity	.400	.443	.352
Old, random	.692	.685	.676
New, random	.237	.232	.233

Table 3
Experiment 2: Best-Fitting Parameters and Summary Fits From the Different Versions of NEMO

	Model Version				
Parameters	A	B	C	D	F
c	2.075	1.962	2.379	2.739	2.762
w_2	0.088	0.086	0.175	0.188	0.175
w_3	0.548	0.561	1.258	1.221	1.247
σ_1	0.432	0.444	0.630	0.618	0.628
σ_2	0.999	0.900	0.839	0.864	0.883
σ_3	0.664	0.667	0.591	0.591	0.593
M_0	1.114	1.000	1.270	3.231	1.198
k	0.218	0.218	0.091	0.137	0.087
β	−1.333	−1.332	0.000	0.000	0.000
Fits					
$-\ln L$	1,399.8	1,422.4	1,631.0	1,498.8	1,620.8
SSD	4.089	4.296	6.119	5.185	6.005
% Var.	83.3	82.5	75.0	78.8	75.5

Note—The values of w_1 and M_1 were held fixed at 1 in all fits. In Version D, $\alpha = 7.492$. In Version F, $v_0 = 0.963$ and $v_1 = 1.013$. (Version E was not fitted, because all the lists were of length 2.)

As is illustrated by the scatterplot in Figure 4, the core model yields a reasonably good account of the complete set of recognition data, although the quantitative fit is not as impressive as that achieved in Experiment 1. Overall, the model accounts for 83.3% of the response variance. In part, the percentage of variance accounted for is somewhat lower in the present experiment, because there is less total variability in the recognition data. A likely reason is that numerous *new* lists in the present experiment were very difficult (i.e., all of the new critical lists in which the probe was adjacent to one of the old study exemplars). Thus, the subjects needed to be conservative in making *old* judgments. Thus, in this experiment, there was a much smaller proportion of lists that had very high or very low recognition probabilities.

We report in Table 2 the predicted recognition probabilities from NEMO (Versions A and C) for the six main list types. This focused comparison makes clear the gains that are yielded by incorporating the β homogeneity parameter. With β held fixed at zero, the summed-similarity exemplar model predicts a much larger false alarm rate for high-homogeneity lists than for low-homogeneity ones (.515 vs. .352), in accord with the intuitions discussed earlier. With β treated as a free parameter, the predicted difference is much smaller (.465 vs. .443) and comes close to matching quantitatively the observed data (.434 vs. .400). A similar pattern is observed for the hit rates associated with the high-homogeneity versus low-homogeneity lists. Thus, the results provide convincing evidence of the need to extend standard summed-similarity exemplar models of *old–new* recognition with parameters related to list homogeneity.

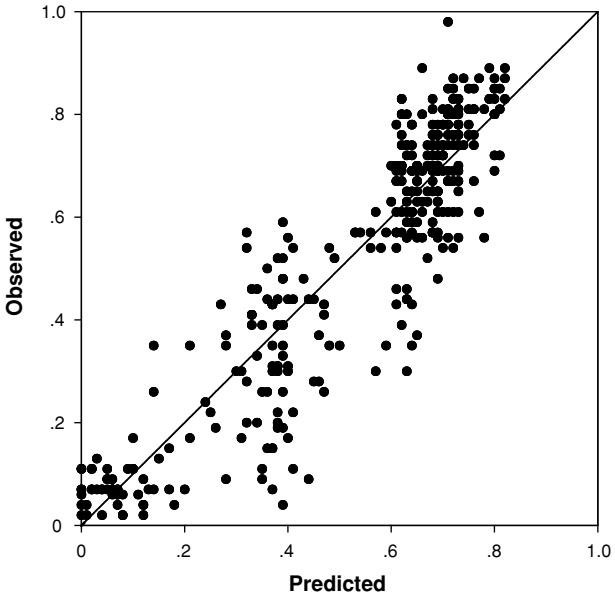

Figure 4. Experiment 2: observed against predicted probabilities of *old* recognition responses for each of the 360 lists.

GENERAL DISCUSSION

The central purpose of the present research was to pursue the highly significant finding of Kahana and Sekuler (2002)—namely, that study list homogeneity exerts an important impact on perceptual *old–new* recognition judgments. Our first concern was that, due to the presence of various emergent dimensions, psychological similarities may not have been measured precisely in the set of stimuli used by Kahana and Sekuler. Thus, the homogeneity parameter could have been correcting for misspecified summed similarities. Second, because of the massive amount of data collected, it is likely that numerous forms of parametric variation would lead to statistically significant improvements in the quantitative fit of models. The question remains about the relative importance of the list homogeneity term, as compared with alternative forms of parametric variation.

Thus, in the present experiments, we conducted a conceptual replication of Kahana and Sekuler's (2002) experiments, except that we used a simpler set of stimuli in which psychological similarities could be more precisely measured. In addition, we conducted extensive comparisons with alternative models in order to test for the relative importance of the role of list homogeneity. In a nutshell, our results provided strong confirmation of Kahana and Sekuler's findings. Indeed, the overall quantitative accounts of the data that were provided by the core model were outstanding and point to a key role of list homogeneity in influencing *old–new* recognition.

Although not discussed in these terms by Kahana and Sekuler (2002), the homogeneity term in the NEMO model can be interpreted in terms of a variable response criterion setting that is dependent on the nature of the study list. According to this interpretation, the observer assesses the overall familiarity of a test item in terms of the standard summed-similarity rule (i.e., Equation 3). If the summed similarity exceeds a criterion, the observer responds *old*; otherwise, the observer responds *new*. However, the criterion that is employed is not fixed but, rather, varies systematically across lists. That is, the observer responds *old* only if the summed similarity S exceeds the criterion $k - \beta \cdot H$. In this sense, the structure of the summed-similarity exemplar model is basically the same as before; it is just that there is systematic variation in the criterion parameter setting across different list types.

Note that in most past tests of exemplar models of recognition, a *single* long list of items would be presented (e.g., Nosofsky, 1991; Shin & Nosofsky, 1992). This single study list was then followed by multiple test items. For each individual test item, an observer would make an *old–new* recognition judgment. Under such conditions involving a single study list, it is natural to assume that parameters remain fixed across the different test items. However, in Kahana and Sekuler's (2002) paradigm, each test item is associated with a separate study list, so parameters might be expected to vary systematically.

Indeed, the present evidence for systematic shifts in response criteria dovetails with recent results reported in the perceptual classification literature. In particular, Cohen, Nosofsky, and Zaki (2001) conducted experiments in which the overall variability of categories was manipulated. Cohen et al. found that observers tended to classify test objects into high-variability categories with higher probability than was predicted by standard summed-similarity exemplar models (for closely related findings, see Rips, 1989; Smith & Sloman, 1994; Stewart & Chater, 2002). They noted that one interpretation of the pattern of results was that the subjects had a response bias (i.e., they set a lower criterion) for classifying objects into high-variability categories than for classifying them into low-variability ones. They also suggested a rational basis for this pattern of behavior. Note that our findings (and those of Kahana and Sekuler, 2002) involving study list homogeneity basically parallel these results involving category variability. That is, low-homogeneity study lists are high-variability ones. Thus, in the same way that subjects are biased to classify objects into high-variability categories, they may be biased to accept test items as old when they have experienced high-variability study lists.

In sum, the results reported in this article, as well as those of Kahana and Sekuler (2002), provide support for the hypothesis that observers make perceptual recognition judgments on the basis of summed similarities of test items to stored exemplars. However, subjects systematically adjust their decision criteria on the basis of the homogeneity of the experienced lists, in a manner parallel to what has been observed in the domain of perceptual classification. A central issue for future research is to understand the psychological basis for such list-specific criterion adjustment.

REFERENCES

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154-179.

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1210-1235.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, **29**, 1165-1175.

Ennis, D. M. (1992). Modeling similarity and identification when there are momentary fluctuations in psychological magnitudes. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 279-298). Hillsdale, NJ: Erlbaum.

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

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

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

Lamberts, K., Brockdorff, N., & Heit, E. (2003). Feature-sampling and random-walk models of individual-stimulus recognition. *Journal of Experimental Psychology: General*, **132**, 351-378.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review*, **79**, 410-419.

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

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology*: *Human Perception & Performance*, **17**, 3-27.

Nosofsky, R. M. (1997). An exemplar-based random walk model of speeded categorization and absolute judgment. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 347-366). Mahwah, NJ: Erlbaum.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). New York: Cambridge University Press.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I. *Psychometrika*, **27**, 125-140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, **121**, 278-304.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, **22**, 377-386.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, **153**, 652-654.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 893-907.

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old–new recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1022-1041.

APPENDIX A
Similarity Scaling of Compound Gratings Stimuli

Here, we will describe the results of the similarity-scaling study that we conducted for the compound gratings stimuli used in the research of Kahana and Sekuler (2002, Experiment 1).

METHOD

Subjects

The subjects were 55 undergraduate and graduate students from Indiana University. Among these subjects, 42 received credit toward an introductory psychology course requirement, whereas 13 were paid $8 for their participation.

Stimuli

The stimuli were the same 27 compound gratings as those used by Kahana and Sekuler (2002) and described in detail in that study. The gratings were presented in pairs in the center of the computer screen against a white background. Each grating filled a 2 × 2 in. square, and the members of the pair were separated by approximately 1 in.

Procedure

The subjects were presented with all 351 distinct pairs of the 27 stimuli. On each trial, they rated the similarity of the members of a given pair on a scale from 1 (*not similar*) to 9 (*very similar*). The order of presentation of the pairs, as well as the left–right placement of the members of each pair, was randomized for each subject.

RESULTS

We analyzed the mean similarity judgments for the 351 pairs of stimuli by using the standard Euclidean model from the ALSCAL program of the SPSS statistical package. We will report here the results of the four-dimensional scaling solution because it yielded good fits to the mean similarity judgments and a systematic correspondence of the derived psychological dimensions with physical aspects of the stimuli. The four-dimensional solution yielded a stress equal to .069 and accounted for 95.9% of the variance in the data.

The four-dimensional scaling solution is displayed graphically in Figure A1. The illustrated solution has been rotated to yield a clear interpretation of the underlying psychological dimensions. The top panel of Figure A1 provides a plot of Dimension 1 against Dimension 2, and the bottom panel provides a plot of Dimension 3 against Dimension 4. The individual stimuli are represented by symbols to bring out the regularity of the derived dimensions. As will be described below, within each plot, all stimuli represented by a common symbol type had an identical value on a given physical dimension. The detailed scaling solution for the individual stimuli is reported in Table A1.

APPENDIX A (Continued)

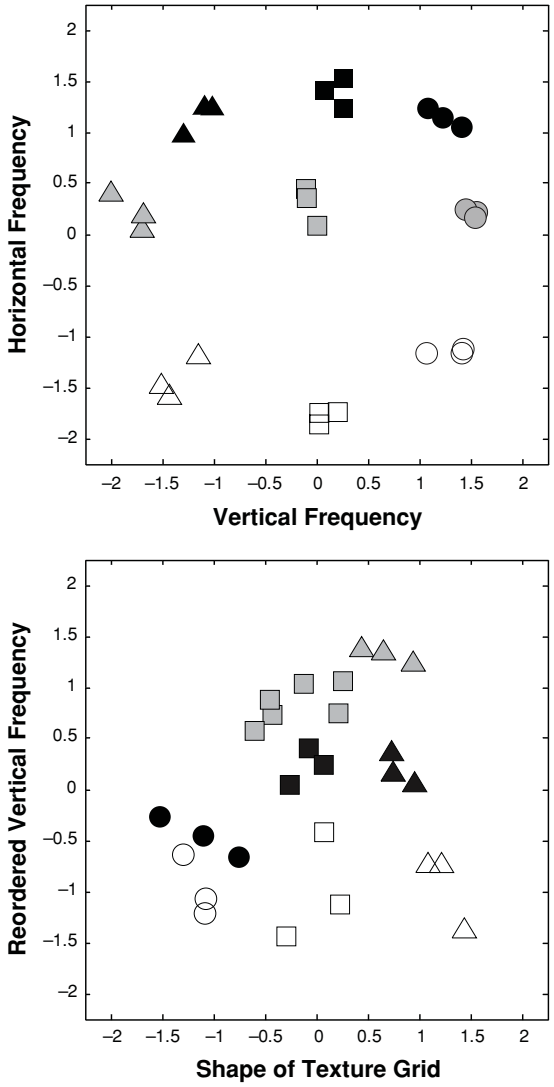

Figure A1. Multidimensional scaling solution for the compound gratings stimuli used in Kahana and Sekuler's (2002) experiment. Top panel: plot of Dimension 1 (vertical frequency) against Dimension 2 (horizontal frequency). Bottom panel: plot of Dimension 3 (shape of texture grid) against Dimension 4 (reordered vertical frequency). See the text for a description of how the different symbol types correspond to variations along the dimensions.

Dimension 1 corresponds clearly to the vertical frequency (V) of each compound grating (triangles, $V = 1$; squares, $V = 2$; circles, $V = 3$). Likewise, Dimension 2 corresponds clearly to the horizontal frequency (H) of each grating (open symbols, $H = 1$; shaded symbols, $H = 2$; solid symbols, $H = 3$). Dimension 3 can be interpreted as the emergent dimension of *shape*. The shape is defined here as the value $|V-H|$. Values of $|V-H| = 0$ correspond to square texture grids and are represented by triangular symbols in the figure; values of $|V-H| = 2$ yield elongated rectangles and are represented by the circular symbols; values of $|V-H| = 1$ yield intermediate rectangles and are represented by the square symbols. Finally, Dimension 4 values correspond to a reordering of the vertical frequency component (open symbols, $V = 1$; solid symbols, $V = 3$; shaded symbols, $V = 2$). Although there is a systematic correspondence of this derived dimension with that of vertical frequency, we are unsure of the psychological interpretation of this reordering.

APPENDIX A (Continued)

Table A1
Multidimensional Scaling Coordinates for the 27 Compound-Gratings Stimuli Used in Kahana and Sekuler's (2002) Experiment

Grating	Physical Coding	Psychological Dimension Values			
		1	2	3	4
1	111	−1.157	−1.201	1.438	−1.394
2	112	−1.502	−1.489	1.219	−0.742
3	113	−1.429	−1.605	1.080	−0.746
4	121	−1.687	0.037	−0.291	−1.445
5	122	−1.682	0.179	0.224	−1.123
6	123	−1.990	0.394	0.076	−0.429
7	131	−1.088	1.245	−1.083	−1.218
8	132	−1.294	0.968	−1.291	−0.635
9	133	−1.025	1.233	−1.066	−1.067
10	211	0.028	−1.759	−0.432	0.726
11	212	0.022	−1.862	−0.453	0.869
12	213	0.199	−1.740	−0.605	0.572
13	221	0.004	0.086	0.945	1.234
14	222	−0.115	0.442	0.655	1.341
15	223	−0.102	0.343	0.440	1.362
16	231	0.272	1.226	−0.127	1.038
17	232	0.078	1.403	0.253	1.065
18	233	0.260	1.514	0.220	0.748
19	311	1.428	−1.119	−1.094	−0.455
20	312	1.412	−1.171	−0.759	−0.666
21	313	1.070	−1.169	−1.522	−0.273
22	321	1.564	0.218	0.067	0.238
23	322	1.455	0.249	−0.067	0.400
24	323	1.546	0.168	−0.260	0.050
25	331	1.089	1.226	0.959	0.052
26	332	1.231	1.141	0.735	0.350
27	333	1.408	1.041	0.737	0.148

Note—Physical coding gives the logical values of each grating along the physical dimensions of vertical frequency, horizontal frequency, and relative phase.

Note that none of the derived psychological dimensions corresponds to the physical dimension of relative phase that was manipulated in Kahana and Sekuler's (2002) experiment. Kahana and Sekuler's application of NEMO yielded an analogous result. Their parameter estimates revealed very low attention weight devoted to phase and very large internal noise estimates associated with phase. Both results indicate that phase entered minimally into the subjects' recognition judgments in their study.

APPENDIX B

Experiments 1 and 2: RGB Values for the Computer-Generated Colors Used in Both Experiments

Stimulus	R	G	B
1	148	169	235
2	144	162	221
3	152	165	207
4	116	137	210
5	113	128	186
6	124	132	172
7	84	107	178
8	91	104	163
9	95	104	147
10	109	173	244
11	123	172	226
12	136	172	209
13	84	146	214
14	94	141	193
15	106	139	178
16	43	110	177
17	60	109	160
18	76	108	146
19	83	177	230
20	102	178	219
21	115	173	230
22	53	150	205
23	74	148	190
24	87	143	173
25	2	113	168
26	37	114	154
27	61	114	144

(Manuscript received June 15, 2004;
revision accepted for publication January 23, 2005.)