Recognizing distinctive faces:
A hybrid-similarity exemplar model account

BETHANY R. KNAPP, ROBERT M. NOSOFSKY, and THOMAS A. BUSEY
Indiana University, Bloomington, Indiana

In recognition memory experiments, Nosofsky and Zaki (2003) found that adding discrete distinctive features to continuous-dimension color stimuli helped participants to identify old items as old (the old-item distinctiveness effect), as well as to identify new items as new. The present study tests the extent to which these results generalize to the domain of face recognition. Two experiments were conducted, one using artificial faces and one using natural faces. Artificial faces were used to test memory for faces with discrete distinctive features while controlling the similarity of the faces themselves on more continuous dimensions. The natural-face experiment used the faces of 40 bald men categorized into three groups (typical, isolated, and distinctive) based on experimental ratings of distinctiveness. In both experiments, there were strong effects of the distinctive features on recognition performance. The data were accounted for reasonably well by a hybrid-similarity version of an exemplar recognition model (Nosofsky & Zaki, 2003), which includes a feature-matching mechanism that can provide boosts to an item's self-similarity.

According to exemplar models of old–new recognition (Gillund & Shiffrin, 1984; Hintzman, 1988; Lamberts, Brockdorff, & Heit, 2003; Medin & Schaffer, 1978; Nosofsky, 1988), people represent lists of study items in terms of individual exemplars, with each exemplar corresponding to an individual study item. In most versions of these models, test items are assumed to give rise to a global activation of the exemplar-based memory representation. The greater the degree of activation, the more familiar is the test item, and the greater is the probability that the observer judges the item to be old.

A representative of this class of global-familiarity exemplar models is the generalized context model (GCM; Nosofsky, 1986, 1991). In the GCM, exemplars are represented as points in a multidimensional psychological space, and the similarity between exemplars is a decreasing function of distance in the space. In the model, global activation or familiarity is based on the summed similarity of a test item to all of the stored study list exemplars. The model has been applied primarily in domains involving the recognition of simple perceptual stimuli that varied along a few salient dimensions, such as colors, schematic faces, and geometric forms varying in size and orientation. In these domains, fine-grained measurements can be obtained of the similarity between exemplars in the multidimensional space. Thus, the model allows fine-grained

predictions of the probability with which individual items are judged as old or new. Indeed, there have been numerous demonstrations of the ability of the model to account in quantitative detail for rich sets of old–new recognition data for individual items (Nosofsky, 1991; Shin & Nosofsky, 1992; Zaki & Nosofsky, 2001; for closely related work, see Kahana & Sekuler, 2002; Lamberts et al., 2003).

However, most of the successes of the GCM have involved its ability to predict how false alarm rates associated with new items vary with their similarity to old items. A potentially important limitation of the standard model is that it fails to account for patterns of results involving hit rates to old items. Specifically, because of its summed similarity rule, the standard GCM predicts that "typical" old items should have higher hit rates than "distinctive" old items (Valentine & Ferrara, 1991). Although the constructs of "typicality" and "distinctiveness" are open-ended, the general idea is that a distinctive item is one that is relatively isolated in the multidimensional similarity space of studied exemplars, whereas typical items lie in more densely clustered locations of the similarity space. Thus, the summed similarity associated with typical items is greater than that associated with distinctive items. Therefore, typical old items should give rise to a greater degree of familiarity, and so should have higher hit rates. A similar qualitative prediction is made by most other global-familiarity models of old–new recognition (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988).

Much research, however, especially from the face recognition literature (Bartlett, Hurry, & Thorley, 1984; Light, Kayra-Stuart, & Hollander, 1979; Valentine & Ferrara, 1991; Vokey & Read, 1992), suggests that distinctive old items have higher hit rates than do typical old items,

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

thereby providing a direct challenge to such models. A good example of such a result is found in a face recognition study conducted by Busey and Tunnicliff (1999). These researchers had participants provide similarity ratings for a large set of naturalistic male faces, analyzed the similarity data by using multidimensional scaling (MDS) techniques, and located the faces in a six-dimensional psychological space. They then used the scaling solution in combination with the GCM to predict participants' performance in an old–new recognition task involving the same set of faces. Busey and Tunnicliff found that distinctive faces—that is, those lying in isolated regions of the multidimensional similarity space—gave rise to *higher* hit rates than did typical faces, in direct contrast to the predictions of the GCM. However, as acknowledged by Busey and Tunnicliff, besides lying in isolated regions of the space, many of the distinctive faces also had highly salient specific features such as beards, whereas faces lying in denser, more typical regions did not. Thus, the isolation of the faces along the continuous dimensions of the space was confounded with the presence of idiosyncratic discrete features.

To investigate further the basis of Busey and Tunnicliff's (1999) results, Zaki and Nosofsky (2001) and Nosofsky and Zaki (2003) explicitly manipulated individual-item distinctiveness in the domain of color. Unlike face space, the underlying dimensional structure of colors is well understood, with extensive scaling work indicating that colors vary along the three dimensions of hue, saturation, and brightness. Thus, one can explicitly manipulate typicality and distinctiveness by varying the location of items along these three dimensions. In several old–new recognition experiments, Zaki and Nosofsky (2001) and Nosofsky and Zaki (2003) tested conditions in which individual colors were located in either densely populated or isolated regions of the continuous-dimension color space. Importantly, across conditions, the same colors served as either typical or distinctive items, thereby removing any stimulus-specific effects from the main pattern of results. When distinctiveness was manipulated in this manner, these researchers found that hit rates for distinctive old items were *not* greater than for typical old items. In fact, the complete set of results was predicted well by the standard GCM. These studies suggested that mere isolation in a continuous-dimension similarity space is not sufficient to produce a robust hit rate advantage for distinctive old items.

Thus, to pursue further the findings of Busey and Tunnicliff (1999), Nosofsky and Zaki (2003) conducted additional experiments in the color domain, in which distinctiveness was manipulated by including idiosyncratic discrete features on the studied objects. Specifically, they tested designs in which various discrete alphanumeric characters were added to a few members of a set of continuous-dimension color patches. Under these conditions, a strong old-item distinctiveness effect was observed: Old items with a discrete distinctive feature added to them had significantly higher hit rates than did old items without such features. This result is similar to the previously discussed results for faces from Busey and Tunnicliff. In addition, as discussed more fully later in this article, adding the distinctive features to foil items made it easier for participants to realize that the foils were new.

To account for the distinctiveness effects, Nosofsky and Zaki (2003) developed a modified version of the GCM called the *hybrid-similarity GCM* (HS-GCM; for closely related ideas in the domains of similarity and classification, see Lee & Navarro, 2002; Navarro & Lee, 2003; Verguts, Ameel, & Storms, 2004). This extended version of the GCM incorporates the ideas of Tversky's (1977) feature contrast model (FCM) into the MDS framework of the traditional GCM. The basic assumption in the FCM is that the similarity between two objects is based on measures of their common and distinctive features. Specifically, similarity is an increasing function of the measure of the objects' common features and a decreasing function of the measures of the objects' distinctive features. The crucially important feature of this model is that, whereas standard MDS approaches assume that all items have an equal degree of self-similarity, the FCM allows for differing degrees of self-similarity, with an increase in the number of common matching features increasing the measure of self-similarity. Thus, within the FCM framework, an item with a highly salient feature will have a higher measure of self-similarity than will one without a highly salient feature. In the HS-GCM, the similarity between two objects is determined jointly by their distance in a continuous-dimension psychological space *and* by the extent to which they have matching or mismatching discrete features. Within this extended framework, if a distinctive item's self-similarity is sufficiently high, its summed similarity can exceed that of more typical items. This increase in summed similarity leads to a corresponding increase in familiarity, and results in an increased probability of judging the distinctive item as old. Indeed, the inclusion of the feature-matching mechanism allowed the HS-GCM to match the quantitative results of Nosofsky and Zaki (2003), whereas the traditional version of GCM could not.

Although the results of Nosofsky and Zaki (2003) provided important preliminary evidence in favor of the new model, the stimuli used in their experiments were highly artificial (color patches combined with alphanumeric markings). Objects in the natural world are rarely that simplistic. Therefore, the main goal of the present research was to explore the extent to which Nosofsky and Zaki's results generalize to a more natural domain—namely, that of face recognition. Given that it has proven difficult for GCM to explain results in this domain in the past (Busey & Tunnicliff, 1999; Valentine & Ferrara, 1991), this test seems especially appropriate for the HS-GCM.

Because faces are highly variable and complex, we conducted two experiments, one using artificial faces and one using natural faces. Artificial faces were used as an experimental intermediary between the easily manipulated domain of color and the much less easily manipulated domain of faces. Artificial faces were derived using principal components analysis on images from the Facial Rec-

ognition Technology (FERET) database (Phillips, Moon, Rizvi, & Rauss, 2000). Because the similarity of these faces on more continuous dimensions could be manipulated, foils of varying degrees of similarity to the distinctive targets could be created (see the Method section), thus allowing for direct comparison of the results with those of Nosofsky and Zaki (2003).

In a second experiment, we provided further tests of the modeling approach in a naturalistic face domain. With natural faces, no predetermined continuous dimensions exist to explicitly create typical versus distinct items. In addition, many potential discrete distinctive features exist on natural faces. To create a reasonable but challenging test bed for the model, we selected a subset of the naturalistic faces used in Busey and Tunnicliff's study. We selected the study and test items so as to vary both item typicality and the number of faces with highly salient discrete distinctive features. A multidimensional face space was constructed by collecting similarity and distinctiveness ratings of the faces used in the experiment. We then used the derived face space and the distinctiveness ratings in combination with the HS-GCM to predict the old–new recognition judgments.

EXPERIMENT 1

The design of Experiment 1 involved a conceptual replication of the earlier study conducted by Nosofsky and Zaki (2003), with the main difference that a set of artificial faces was used instead of a set of color patches. In addition, whereas Nosofsky and Zaki used arbitrary alphanumeric characters as discrete distinctive features, in the present experiment we used distinctive features that are plausibly viewed on faces, such as scars, glasses, and so forth. Because the psychological space in which faces are embedded is composed of highly complex and unspecified dimensions, we could no longer manipulate continuous-dimension similarity in the same manner as in the Nosofsky and Zaki study. However, we used an alternative method that created compact clusters of faces, with each cluster lying in a separate region of the face space. The key question was the extent to which the pattern of results from Nosofsky and Zaki's design with highly artificial color/alphanumeric-character stimuli would generalize to this richer, more complex, and more naturalistic face domain.

Method

Participants

The participants were 121 undergraduates at Indiana University who received partial credit toward an introductory psychology course, as well as the added incentive of a $10 bonus for the top three performers in the experiment. The participants were tested in groups of 4 to 6.

Stimuli

A total of 72 artificial faces were used in the experiment. The faces were constructed by using procedures adapted from O'Toole, Abdi, Deffenbacher, and Valentin (1993). First, we aligned 500 faces from the FERET database (Phillips et al., 2000) by scaling and ro-

tating each face such that the two eyes and mouth were located in common locations, taken from the averages across all faces. The faces were then submitted to a principal components analysis that preserved 20 eigenvectors. Such eigenvectors are interpreted as patterns of variation that exist in the set of faces. The original faces can be reconstructed, with various amounts of error, by weighted linear combinations of these eigenvectors. To create clusters of faces, we chose reconstructed faces that used the weights corresponding to existing faces. The 20 weights placed each face in a 20-dimensional space, and we termed these reconstructed faces the *prototype* faces. A cluster of faces was generated by taking a vector of fixed length and rotating it randomly in 20-dimensional space. The tail was placed at the coordinates of the prototype face, and the head of the vector defined a new face. This procedure was repeated 20 times to produce a cluster of faces that had the property that each face was approximately equally distant from the prototype. We generated 200 such clusters and then hand-selected a subset of the constructed stimuli for use in the experiment.

The design of the stimulus set is illustrated schematically in Figure 1. Twelve clusters of faces were selected. Within each cluster, there were six similar faces that were roughly equidistant from the prototype. The prototypes defining each cluster were quite distant from one another. Thus, faces within each cluster were similar, whereas faces between clusters were dissimilar. Because we judged the prototypes as looking significantly more like "real" faces than did the rotated faces, the prototypes were not used in the experiment. Three faces from each cluster were randomly assigned to be old items (seen during both study and test). The remaining faces in each cluster, the *foils*, were seen only at test. Six clusters were randomly assigned to be "marked." Within each marked cluster, a discrete distinctive feature was added to one of the old faces. Each of these faces (called a *distinctive old item*) received one of the following discrete distinctive features, with each feature being used on exactly one face during study: glasses, headband, mole, scar, forehead wrinkles, and deep smile lines around the mouth. During the test phase, each discrete feature was also presented on one of the three foils from each distinctive target's cluster. (The same feature was used on targets and foils from the same cluster.) We refer to the latter items as the *high-similarity distinctive foils*, because they are highly similar to the distinctive target from their own cluster. In addition, each discrete feature was also placed on one of the three foils from each of the six unmarked clusters. We refer to these items as the *low-similarity distinctive foils*, because they came from a cluster separate from the distinctive target. Thus, three faces (one distinctive old, one high-similarity distinctive foil, and one low-similarity distinctive foil) had each of the six distinctive features at test. The remaining study items (called *typical old items*) and foil items (called *typical foils*) were left unmarked, and these items constituted the majority of the stimulus set. In total, there were 36 old items (30 typical old items and 6 distinctive old items) and 36 new items (24 typical foils, 6 high-similarity distinctive foils, and 6 low-similarity distinctive foils). An example of the actual stimulus materials for one of the clusters of faces is shown in Figure 2.

Procedure

Stimulus assignment was randomized for each group of participants; specifically, we randomized which clusters were marked versus unmarked and which items within clusters were assigned as typical old items, distinctive old items, typical foils, high-similarity distinctive foils, and low-similarity distinctive foils. During study, the participants were shown all of the old items in two blocks, with each face shown in a random order once per block. The presentation time for each item was 2 sec. The test phase followed immediately after the study phase: All 72 items were presented individually in random order, and the participants were asked to indicate whether each face was old or new by pressing an appropriate key on a keypad. Each face remained in view until all of the participants had responded. No feedback was provided during the test phase.
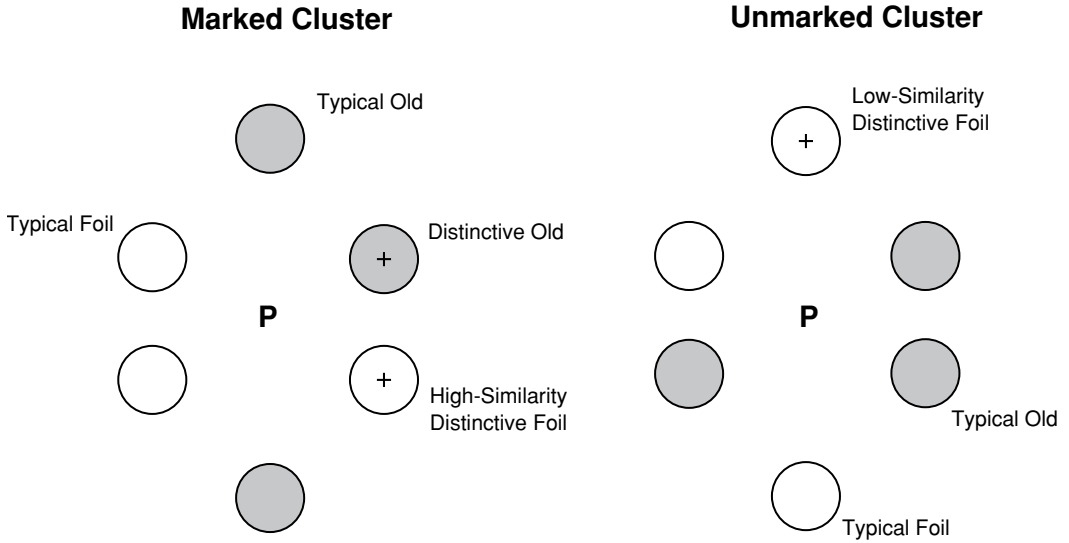
Marked Cluster   **Unmarked Cluster**

Figure 1. Schematic design of Experiment 1 showing one "marked" and one "unmarked" cluster. All faces within a cluster are statistical distortions of a central prototype (P). Shaded symbols represent old study items, and unshaded symbols represent foils. The "+" sign represents a discrete distinctive feature. Each marked cluster had one old study item with a discrete distinctive feature (*distinctive old item*) and one foil with the same discrete distinctive feature (*high-similarity distinctive foil*). Each unmarked cluster included one foil with a discrete distinctive feature (*low-similarity distinctive foil*). Old study items without discrete distinctive features are termed *typical old items*, whereas foils without discrete distinctive features are termed *typical foils*.

Results

The mean probability of *old* judgments for each item type is reported in Table 1. Not surprisingly, the hit rate for typical old items ($M = .55$) was significantly higher than the false alarm rate for typical foils ($M = .44$) [$t(120) = 7.67, p < .001$], reflecting the participants' ability to discriminate old from new items. Note, however, that this difference between hit and false alarm rates is not very large,

so old–new discrimination was difficult for the present set of artificial faces. The most important result is that an old-item distinctiveness effect was observed, with the hit rates of distinctive old items ($M = .73$) being significantly higher than those for typical old items ($M = .55$) [$t(120) = 8.56, p < .001$]. As was observed by Nosofsky and Zaki (2003), the false alarm rate for typical foils ($M = .44$) was significantly higher than that for low-similarity
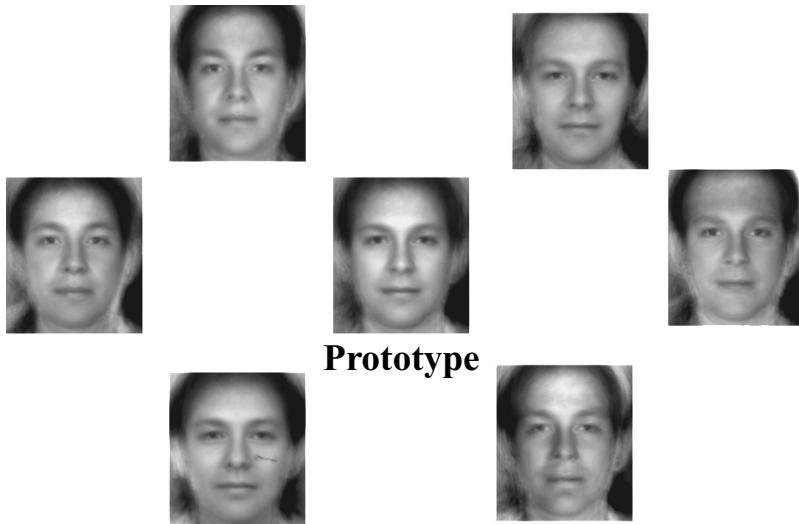

Figure 2. A sample cluster of the artificial faces used in Experiment 1, with the prototype face in the center of the cluster. The face at the lower left of the figure illustrates an example in which a distinctive feature has been added (in this case a scar).

Table 1
Observed and Predicted *Old* Recognition Probabilities
for the Seven Item Types in Experiment 1

Region and Item Type	Observed	HS-GCM	GCM
Marked clusters			
Distinctive old	.73	.73	.57
Typical old	.54	.54	.58
Typical foil	.44	.43	.43
High-similarity distinctive foil	.50	.50	.41
Unmarked clusters			
Typical old	.56	.56	.59
Typical foil	.45	.45	.45
Low-similarity distinctive foil	.34	.34	.39

Note—HS-GCM, hybrid-similarity generalized context model; GCM, generalized context model.

distinctive foils ($M = .34$) [$t(120) = 7.67, p < .001$]. However, unlike Nosofsky and Zaki (2003), the false alarm rate for high-similarity distinctive foils ($M = .50$) was significantly higher than that for typical foils ($M = .44$) [$t(120) = 2.84, p = .01$], rather than lower. We consider these results more fully in the Theoretical Analysis section below.

Theoretical Analysis

According to the HS-GCM, the probability of an *old* response is defined in terms of global familiarity (F_i). The probability that item i is judged as old is given by

$$P(\text{old} \mid i) = \frac{F_i}{F_i + k}, \qquad (1)$$

where k is a response criterion parameter. The global familiarity of item i is measured by the summed similarity of i to each study exemplar j:

$$F_i = \sum_j s_{\text{H}}(i, j), \qquad (2)$$

where $s_{\text{H}}(i, j)$ is the hybrid similarity of item i to exemplar j. The measure of similarity in the hybrid model extends the traditional GCM measure by incorporating two new parameters, C and D, which take into account the presence of the discrete distinctive features. In particular, similarity in the hybrid model is defined as

$$s_{\text{H}}(i, j) = C \cdot D \cdot s(i, j), \qquad (3)$$

where $C > 1$ is a common-feature match parameter that measures the boost in similarity resulting from matching discrete features; D ($0 < D < 1$) is a distinctive-feature mismatch parameter that measures the reduction in similarity resulting from mismatching discrete features; and $s(i, j)$ is the traditional GCM measure of the similarity of items i and j in a continuous-dimension space.

In numerous applications of the GCM, continuous-dimension scaling solutions are first derived in order to compute interexemplar similarities. For simplicity, however, in the present case we approximate the measure of continuous-dimension similarity in terms of two additional parameters: s_{w}, the mean similarity of the faces within a cluster of faces; and s_{b}, the mean similarity among

faces across different clusters. Finally, as in the traditional GCM, the continuous-dimension similarity of an item to itself is always equal to 1. Thus, the familiarity for each type of item was defined in terms of the parameters C, D, s_{w}, and s_{b}. A similar approach was used by Nosofsky and Zaki (2003). So, for example, for the present design the summed similarity for distinctive old items would be given by

$$F_i = C + (2 \cdot s_{\text{w}} \cdot D) + (33 \cdot s_{\text{b}} \cdot D). \qquad (4)$$

The term C measures the distinctive old item's self-similarity; the term ($2 \cdot s_{\text{w}} \cdot D$) measures the summed similarity of the distinctive old item to the two typical old items in its cluster; and the term ($33 \cdot s_{\text{b}} \cdot D$) measures the summed similarity of the distinctive old item to the 33 old items from the other 11 clusters. To take another example, the summed similarity for high-similarity distinctive foils would be given by

$$F_i = (s_{\text{w}} \cdot C) + (2 \cdot s_{\text{w}} \cdot D) + (33 \cdot s_{\text{b}} \cdot D). \qquad (5)$$

In this case, the term ($s_{\text{w}} \cdot C$) measures the high-similarity distinctive foil's similarity to the distinctive old item from within its own cluster; the term ($2 \cdot s_{\text{w}} \cdot D$) measures the summed similarity of the foil to the two typical old items in its cluster; and the term ($33 \cdot s_{\text{b}} \cdot D$) measures the summed similarity of the foil to the 33 old items from the other 11 clusters.[1] Similar equations can be derived for each of the other five item types. Note that the present application of HS-GCM involves the use of five free parameters: s_{w}, s_{b}, C, D, and k.

We conducted a computer search for the values of the five free parameters that provided a maximum-likelihood fit to the old–new recognition data. Specifically, we searched for the free parameters that maximized the log-likelihood function

$$\ln(L) = \sum \ln(N_i!) - \sum\sum \ln(f_{ij}!) + \sum\sum f_{ij}\ln(p_{ij}), \qquad (6)$$

where N_i is the frequency with which stimulus type i is presented; f_{ij} is the frequency with which participants judge item type i as old ($j = 1$) or new ($j = 2$); and p_{ij} is the predicted probability with which item type i is judged as old or new. (This likelihood function assumes that the *old* and *new* judgments for each item type are binomially distributed and that the individual responses are independent.)

To help evaluate the fits of alternative models with differing numbers of free parameters, we used the Bayesian information criterion (BIC; Wasserman, 2000). The BIC fit for a model is given by

$$\text{BIC} = -2\ln(L) + P \cdot \ln(N), \qquad (7)$$

where $\ln(L)$ is the log-likelihood of the data, P is the number of free parameters in the model, and N is the total number of observations in the data set. The model that yields a smaller BIC value is the preferred model. In the BIC, the term $P \cdot \ln(N)$ penalizes a model for its number of free parameters. Thus, if two models yield nearly equivalent log-likelihood fits to the data, then the simpler model with fewer free parameters is preferred.

The predicted probabilities from the full version of the HS-GCM are shown with the observed probabilities in Table 1. The best-fitting values of the free parameters are reported in Table 2. The model accounts for 99.9% of the variance in the *old* recognition probabilities and yielded BIC = 98.1. As can be seen from Table 1, the model captures well the magnitude of the old-item distinctiveness effect. Furthermore, an important role of self-match was indicated in this experiment, since the common-feature parameter took the value $C = 4.12$. There was also a relatively large reduction in similarity due to mismatching distinctive features, with $D = .56$. Thus, besides providing excellent quantitative fits to the data, the resulting parameter estimates from the model vary in highly interpretable ways.

It is important to note that whereas Nosofsky and Zaki (2003) found that the false alarm rate for high-similarity distinctive foils was significantly lower than for typical foils, the opposite pattern was observed in the present experiment (see Table 1). Furthermore, the HS-GCM accounts for both patterns of findings. The explanation in terms of the model is as follows: When a distinctive discrete feature is added to a foil, it has two competing influences. First, it results in a boost to overall summed similarity, because the high-similarity distinctive foil matches the distinctive target on this salient discrete feature. This boost is captured by the C parameter in the term $(s_\mathrm{w} \cdot C)$ in Equation 5. Second, it results in a reduction to overall summed similarity, because that foil now has a greater mismatch to all of the other old targets in the study list. This reduction is captured by the D parameter in the terms $(2 \cdot s_\mathrm{w} \cdot D) + (33 \cdot s_\mathrm{b} \cdot D)$ in Equation 5. The effect on the false alarm rate thus depends on the outcome of this competition. In Nosofsky and Zaki's (2003) experiment with color patch stimuli, within-cluster similarity (s_w) was relatively low, so the boost to summed similarity ($s_\mathrm{w} \cdot C$) did not overcome the reduction. By contrast, in the present experiment, within-cluster similarity was relatively high, which is reflected by the difficulty that participants had in discriminating new foils from old targets (compare typical foils vs. typical old items in Table 1). Thus, the boost to summed similarity greatly outweighed the reduction, and there was a corresponding increase in false alarm rates. As impressive further support for the model, note that it simultaneously predicts the boost in false alarm rates for the high-similarity distinctive foils, *as well as* the reduction in false alarm rates for the low-similarity distinctive foils. In a nutshell, the effect of adding distinctive discrete features on false alarm rates depends on parametric variations in similarity within the stimulus set, and these effects are well captured by the HS-GCM.

Finally, as a source of comparison, we also fitted to the old–new recognition data a special case of the HS-GCM in which the common-feature match parameter was held fixed at $C = 1$. This special-case model is essentially the standard GCM—in other words, a version of the exemplar model in which all self-similarities are equal and held fixed at 1. The predicted probabilities from the standard GCM are shown in Table 1, with the best-fitting parameters reported in Table 2. As can be seen, the standard model fails dramatically to predict the old-item distinctiveness effect. It also fails to predict the increase in false alarms associated with the high-similarity distinctive foils. The standard model yielded BIC = 232.3, which is a considerably worse fit than was yielded by the full HS-GCM (BIC = 98.1). Thus, the extended model captures fundamental psychological phenomena that the standard GCM fails to capture.

Discussion

The results from Experiment 1 provide a conceptual replication of the earlier results reported by Nosofsky and Zaki (2003). When distinctive discrete features are added to individual stimuli embedded in a continuous-dimension similarity space, a robust old-item distinctiveness effect is observed. The standard GCM fails to account for the effect, whereas an extended hybrid-similarity version that makes allowance for the role of discrete-feature matching performs extremely well. Furthermore, beyond accounting for the old-item distinctiveness effect, the hybrid-similarity model also captures the effect of the discrete features on false alarm rates for both high-similarity and low-similarity foils. Finally, the model performs well in a fairly complex and naturalistic domain—namely, that of face recognition.

One question that arises is whether or not the results can be explained in terms of differential encoding and memory strengths of the distinctive items rather than in terms of differences in self-similarity. For example, during study, the distinctive items may have captured the observers' attention, resulting in stronger memory traces in comparison with the typical old items (cf. Malmberg & Nelson, 2003). The previous work conducted by Nosofsky and Zaki (2003), however, argued against such an interpretation. In that work, conditions were included in which the frequency with which the distinctive features appeared on foil items at the time of test was varied. In conditions in which the distinctive features rarely appeared on the

Table 2
Best-Fitting Parameters for the HS-GCM and GCM in Experiments 1 and 2

Parameter	Experiment 1		Experiment 2	
	HS-GCM	GCM	HS-GCM	GCM
C	4.12	1.00 (set)	–	–
D	.56	.81	–	–
s_w	.27	.30	–	–
s_b	.02	.00	–	–
α	–	–	.20	–
β	–	–	.12	–
γ	–	–	.14	–
k	1.75	1.12	0.56	0.22
κ	–	–	1.24	2.71

Note—HS-GCM, hybrid-similarity generalized context model; GCM, generalized context model; C, common discrete feature match; D, distinctive discrete feature mismatch; s_w, mean similarity of the faces within a cluster of faces; s_b, mean similarity among faces in different clusters of faces; α, β, γ, distinctive-feature weight parameters; k, response criterion; κ, overall sensitivity.

foils, the presence of a distinctive feature was highly diagnostic that the test item was old. By contrast, in conditions in which the distinctive features occurred with high frequency on the foils, the presence of a distinctive feature provided weak evidence that the test item was old. Observers were apparently sensitive to this manipulation, because the common-feature match parameter took on very high values in the conditions in which the distinctive features were diagnostic, but very low values when the distinctive features were not diagnostic. The memory strength hypothesis fails to explain this pattern of results, because any differential encoding of the distinctive items occurs at time of study, prior to the presentation of the test items. By contrast, the pattern of results is explained by the hypothesis that at time of test, participants adjust the weight assigned to common-feature matches in accord with the judged diagnosticity of the discrete distinctive features.

EXPERIMENT 2

The HS-GCM accounts well for the recognition data from artificial faces, in which the typicality of the faces within and between clusters could be experimentally controlled. Yet, though the artificial faces were far more complex than the color patches used by Nosofsky and Zaki (2003), they still lacked many of the properties that natural faces possess. Most critically, in our second experiment we were interested in testing for a role of discrete-feature structure as it might occur in the natural world, rather than in explicitly introducing such structure artificially. Therefore, our next step was to explore the extent to which our previous results generalize to the more complex domain of natural faces.

As a challenging test bed for the model, we used as stimuli a 40-face subset of the 104 naturalistic faces that had been used by Busey and Tunnicliff (1999). (Recall that the predictions from the traditional GCM had been strongly contradicted in this previous study.) We tested three conditions in the present experiment, all involving the same set of naturalistic faces. In one condition, we obtained similarity judgments of the faces to generate an MDS solution. In a second condition, a separate group of participants engaged in an old–new recognition task. Following this task, we tested a third condition in which the participants provided individual-feature distinctiveness ratings of the faces. These ratings were used to obtain information concerning the discrete-feature structure associated with the set of faces. The goal was to use the derived MDS solution together with the distinctiveness ratings to test the HS-GCM on its ability to predict the old–new recognition judgments.

We chose the subset of 40 faces with the aim of tapping into both the continuous-dimension and discrete-feature components of similarity of the HS-GCM. One group of faces was chosen with the expectation that the members would cluster together in the MDS solution, making them *typical* items. Another group was chosen with the expectation that its members would lie in more isolated portions

of the MDS solution, but such that the faces did not have any highly salient discrete features. We refer to this group as the *isolated* faces. Finally, a third group was chosen such that its members were not only expected to lie in isolated portions of the MDS solution, but also possessed unique and highly salient discrete features. We refer to this group as the *distinctive* faces.

The central prediction from the HS-GCM is that there will be marked increases in hit rates for the old distinctive faces as a result of feature matching, whereas foils with the discrete features will show marked decreases in false alarm rates because of feature mismatching. It is more difficult to make a priori predictions comparing hit rates of the typical old and isolated old faces. On the one hand, because typical faces lie in dense regions of the similarity space, they should give rise to higher levels of summed similarity. On the other hand, in this naturalistic domain, it is impossible to ensure that observers will not notice any idiosyncratic discrete features on the isolated faces, despite the experimenters' judgments that none were highly salient. And to the extent that such idiosyncratic discrete features are more likely to exist on the isolated faces than on the typical ones, hit rates for the isolated faces may be slightly elevated as well.

Method

Participants

The participants were 159 undergraduates at Indiana University taking introductory psychology courses. They received course credit for their participation, and an additional $3 performance bonus was also offered. There were 72 participants in the similarity ratings condition and 87 in the old–new recognition and distinctiveness ratings conditions. Because of poor performance on the old–new recognition task, the data of 4 participants were not included in the analyses. All participants were tested individually.

Stimuli

Busey and Tunnicliff (1999) conducted experiments using pictures of 104 bald men (Kayser, 1985) with similar expressions, taken under similar lighting conditions. The 104 pictures included men of various races as well as morphs of various pairs of faces. In choosing the present 40-face subset from Busey and Tunnicliff's stimuli, we used only Caucasians. In addition, because morph faces have been shown to have artifacts that may influence recognition judgments (see, e.g., Busey, 1998; Zaki & Nosofsky, 2001), the morphs were not included in the set.

As described previously, we chose the 40 faces with the aim of creating three main groups: typical, isolated, and distinctive. The results from the distinctiveness-ratings condition were used to confirm our judgments about the group assignments. There were a total of 16 typical faces, 16 isolated faces, and 8 distinctive faces.[2] Half of the faces in each group were randomly designated as old items, with the other half serving as foils during test. Because our goal in this experiment was the modeling of individual-item performance (see the Theoretical Analysis section), the old/foil designation of each face was held constant across participants. Thus, in total there were 20 old items (8 typical old, 8 isolated old, and 4 distinctive old) and 20 new items (8 typical foils, 8 isolated foils, and 4 distinctive foils). Regarding the distinctive-face groups, for purposes of simplicity in our subsequent modeling, we tried to select items in which a different type of discrete feature was associated with each face. The resulting salient discrete features were bushy beard, long curled mustache, trimmed gray goatee, missing eyebrows, highly wrinkled face, extremely round face, forehead scar, and prominent cheeks.[3]

Procedure

Similarity judgment condition. Each participant was shown half of the 780 possible pairs of the 40 faces in a side-by-side presentation and asked to judge how similar the two faces were on a scale of 1 (*not very similar*) to 9 (*very similar*). Responses were made by clicking on the appropriate rating using a mouse. After rating the current pair of faces, the participants could bring up a new pair of faces by pressing a Finished button with the mouse. The pairs of faces judged by each participant were chosen randomly. The left–right placement of the faces on the screen was also chosen randomly.

Old–new recognition condition. During study, these participants were shown all 20 old faces in a random order in two blocks, such that each face was shown once per block. The presentation time for each face was 4 sec. The test phase immediately followed study. During the test phase, an individual face was presented for 4 sec, followed by a response screen on which the participants were asked to indicate whether that face was old or new. They also indicated how confident they were about their judgment on a scale of 1 (*not very confident*) to 9 (*very confident*). The test faces were presented in a random order for each individual participant.

Distinctiveness ratings condition. Following the old–new recognition task, the participants were asked to rate the extent to which each of the 40 faces had highly distinctive discrete features. The faces were presented in a random order for each participant. After an individual face was presented for 4 sec, the participants were asked to give a rating on a scale of 1 (*not very distinctive features*) to 9 (*very distinctive features*) by clicking on the appropriate response with the mouse. In addition to rating each face, they were also asked to indicate which particular feature or features were distinctive. They could indicate between 0 and 5 features using a drop-down menu that listed 16 features (see Table 3), and they could also choose the "Other" category, which allowed them to write in their own feature. The 16 listed features were compiled from the responses given during a previous unpublished experiment from Busey and Tunnicliff (1999) that used all 104 faces.

Results

Similarity Ratings

Each pairing of faces was rated by an average of 35.5 participants. We analyzed the matrix of averaged ratings by using the standard Euclidean model from the ALSCAL statistical package. A six-dimensional MDS solution had STRESS = .096 and accounted for 86.9% of the variance in the averaged ratings. We used this MDS solution in combination with the HS-GCM in our subsequent modeling analyses. The raw similarity ratings and MDS solution for the faces are available from the authors upon request.

In general, the typical faces were located in central portions of the MDS solution, the isolated faces were located in intermediate regions, and the distinctive faces were located

Table 3
The 16 Features Provided to Participants During the Rating Portion of Experiment 2

Feature	Feature
Beard	Chin
Mustache	Ears
Nose	Face shape
Mouth	Head shape
Lips	Forehead
Eyes	Wrinkles
Eyebrows	Scar
Cheeks	Mole

toward the periphery. The average distance of each type of face to other members of the set was 3.16, 3.38, and 3.83, for the typical, isolated, and distinctive faces, respectively.

Distinctiveness Ratings

The mean distinctiveness rating for each individual face is reported in Table 4, with the mean for each item type reported in Table 5. Mean distinctiveness ratings were significantly higher for distinctive old items ($M = 7.59$) than for isolated old items ($M = 5.99$) [$t(82) = 18.47, p < .001$] and for isolated old items than for typical old items ($M = 4.96$) [$t(82) = 11.29, p < .001$]. Similarly, mean distinctiveness ratings for distinctive foils ($M = 7.78$) were significantly higher than those for isolated foils ($M = 6.03$) [$t(82) = 13.64, p < .001$], and distinctiveness ratings for isolated foils were significantly higher than those for typical foils ($M = 4.66$) [$t(82) = 14.21, p < .001$]. This pattern of results provides a manipulation check on the groups of faces that we formed. The faces that we judged to have the most salient distinctive features (groups distinctive old and distinctive foil) were given the highest distinctiveness ratings by the experimental participants. Note, however, that the isolated faces were also judged to have more distinctive individual features than did the typical faces, although the magnitude of these ratings was not nearly as high as it was for the distinctive-face group. As discussed earlier, this result was not entirely unexpected, given the complexity of the natural-face domain, and needed to be incorporated in our subsequent modeling of the old–new recognition data.

Old–New Recognition

The observed proportion of *old* recognition judgments for each individual face is reported in Table 4. The mean probabilities of *old* judgments for the main item types are reported in Table 5. The mean recognition probability for the distinctive old items ($M = .89$) was significantly higher than for either the isolated old items ($M = .82$) [$t(82) = 3.16, p = .002$] or the typical old items ($M = .82$) [$t(82) = 3.46, p = .001$]. Recognition probabilities for the typical old and isolated old items did not differ significantly [$t(82) = 0.235, p = .82$]. The false alarm rate for distinctive foils ($M = .01$) was significantly lower than for either typical foils ($M = .17$) [$t(82) = 8.76, p < .001$] or isolated foils ($M = .10$) [$t(82) = 6.06, p < .001$]. In addition, the false alarm rate for isolated foils was significantly lower than for typical foils [$t(82) = 4.25, p < .001$]. The increased hit rate associated with the distinctive old items is consistent with the general predictions from the HS-GCM. In addition, the decreased false alarm rates associated with the isolated foils and distinctive foils are straightforward qualitative predictions from the model. We now turn to the formal modeling analyses to test the extent to which the HS-GCM can account quantitatively for the details of the old–new recognition data.

Theoretical Analysis

Our central goal was to use the HS-GCM to predict quantitatively the recognition probabilities associated

Table 4
Observed and Predicted *Old* Recognition Probabilities and
Distinctiveness Ratings for the 40 Faces in Experiment 2

Face Type	Face	Observed Probability	HS-GCM	GCM	Distinctiveness Rating
Typical old	1	.93	.80	.83	4.49
	2	.90	.81	.82	4.73
	3	.86	.81	.83	4.57
	4	.86	.82	.82	5.37
	5	.84	.81	.82	4.77
	6	.83	.81	.82	4.88
	7	.76	.82	.82	5.41
	8	.58	.82	.82	5.42
Typical foil	1	.30	.31	.44	4.19
	2	.30	.15	.16	5.45
	3	.22	.18	.25	3.82
	4	.18	.16	.14	4.96
	5	.18	.15	.10	4.41
	6	.07	.15	.06	4.27
	7	.06	.14	.18	5.27
	8	.06	.06	.02	4.92
Isolated old	1	.95	.85	.82	6.40
	2	.89	.85	.82	6.41
	3	.88	.85	.82	6.49
	4	.83	.85	.82	6.65
	5	.82	.83	.83	5.51
	6	.78	.82	.82	5.41
	7	.72	.83	.82	5.55
	8	.71	.83	.82	5.48
Isolated foil	1	.23	.30	.42	5.07
	2	.23	.10	.08	6.34
	3	.10	.04	.04	5.95
	4	.06	.09	.10	6.25
	5	.06	.08	.04	6.16
	6	.06	.05	.02	5.86
	7	.05	.14	.18	6.07
	8	.05	.08	.12	6.51
Distinctive old	1	.95	.87	.82	7.64
	2	.92	.86	.82	7.23
	3	.86	.87	.82	7.40
	4	.83	.88	.82	8.08
Distinctive foil	1	.02	.03	.07	8.42
	2	.01	.02	.03	7.84
	3	.01	.02	.00	8.01
	4	.00	.03	.06	6.84

Note—HS-GCM, hybrid-similarity generalized context model; GCM, generalized context model.

with each of the individual faces. We should make clear at the outset that we view this goal of predicting recognition probabilities at the individual-item level to be a highly ambitious one in this natural-face domain. Although exemplar recognition models have performed admirably in more simplified domains in which the underlying dimensional structure of the objects is tightly controlled, such control is not possible in the present domain. Therefore, it is unrealistic to expect the model to perform with the same precision as in previous tests.

Note as well that we are unable to provide a rigorous application of the model, because scaling techniques are not yet available for cleanly identifying the continuous-dimension and discrete-feature components of these complex natural-face representations. (As described in our General Discussion section, however, Navarro and

Lee [2003] have achieved some promising developments along these lines.) Nevertheless, our idea was to use the obtained similarity and distinctiveness ratings as vehicles for approximating the model's predictions. Specifically, we used the derived MDS solution for the faces to approximate the continuous-dimension distance component in the model. Furthermore, as will be seen, as an approximation of incorporating the discrete-feature structure of the faces, we made use of the individual-feature distinctiveness ratings provided by the participants.

To begin, the continuous-dimension Euclidean distance $d(i, j)$ between each pair of faces i and j was computed from the six-dimensional scaling solution that we derived from the similarity judgments. The hybrid distance $\tilde{d}(i, j)$ between each pair of faces was then computed by combining the continuous-distance measure with information

Table 5
Average Observed and Predicted *Old* Recognition Probabilities
and Observed Distinctiveness Ratings
for the Six Item Types in Experiment 2

Item Type	Observed Probability	HS-GCM	GCM	Distinctiveness Rating
Typical old	.82	.81	.82	4.96
Typical foil	.17	.16	.17	4.66
Isolated old	.82	.84	.82	5.99
Isolated foil	.10	.11	.12	6.03
Distinctive old	.89	.87	.82	7.59
Distinctive foil	.01	.02	.04	7.78

Note—HS-GCM, hybrid-similarity generalized context model; GCM, generalized context model.

provided from the individual-feature distinctiveness ratings. Specifically, the hybrid distance between two distinct faces i and j was defined as

$$\tilde{d}(i,j) = d(i,j) + \alpha \cdot (D_i + D_j) - \beta \cdot \delta(i,j) \cdot (D_i + D_j), \quad (8)$$

where $d(i,j)$ is the continuous-dimension component; D_i is the distinctiveness rating for face i; and α and β are freely estimated weight parameters that reflect the relative contributions of the discrete distinctive features to the total distance calculation. Finally, $\delta(i,j)$ is an indicator variable set equal to 1 when the most salient rated feature of face i is the same as the most salient feature of face j, and set equal to 0 otherwise. The basic assumption underlying Equation 8 is that the distinctiveness ratings for the individual discrete features are proportional to the features' true "psychological salience." These salience values add to the total hybrid distance between the two faces [i.e., the term $\alpha(D_i + D_j)$]. The final term $[\beta\delta(i,j)(D_i + D_j)]$ is intended as a correction to this added psychological distance: If two faces share a similar salient discrete feature, the added distance is not as great as when the discrete features are quite different from one another.

As is assumed in the traditional GCM, the hybrid similarity between faces i and j is assumed to be an exponential decay function of their distance,

$$s_H(i,j) = \exp[-\kappa \cdot \tilde{d}(i,j)], \quad (9)$$

where κ is an overall sensitivity parameter that measures the rate at which similarity declines with increasing distance. In addition, the similarity of a face to itself is defined as

$$s_H(i,j) = \exp(\gamma \cdot D_i), \quad (10)$$

where $\gamma > 0$ is a freely estimated weight parameter that measures the importance of common-feature matches.

Note that Equations 8–10 yield the same functional form for defining hybrid similarity as was introduced in Experiment 1 (Equation 3), except that allowance is now made for the idea that different discrete features may have differing degrees of psychological salience.[4] In all other respects, the HS-GCM is the same as the version discussed earlier in this article, with the overall familiarity of each face defined by Equation 2, and the *old* recognition probability for each face given by Equation 1.

In the present context, the HS-GCM has five free parameters: the overall sensitivity parameter κ, response criterion parameter k, and hybrid-distance weight parameters α, β, and γ. In the special case in which the hybrid-distance weight parameters are set equal to 0, the model reduces to the traditional GCM. We conducted a computer search for the values of the free parameters that provided a maximum-likelihood fit to the individual-face recognition data and used the BIC statistic to evaluate alternative versions of the model with differing numbers of free parameters.

The predicted recognition probabilities for each individual face from the full version of the HS-GCM are reported along with the observed data in Table 4, with the best-fitting parameters reported in Table 2. The model provides a fair fit to the individual-face data, accounting for 96.0% of the response variance in the *old* recognition probabilities and yielding BIC = 335.15. The predictions of the main trends are reported along with the observed data in Table 5. As can be seen, the model accounts reasonably well for the magnitude of the old-item distinctiveness effect (i.e., the advantage in hit rates for the distinctive old items in comparison with the isolated old and typical old items). Likewise, the model predicts extremely accurately the decrease in false alarm rates as the discrete features on the foils become more salient.

Nevertheless, despite accounting for the main trends, it is important to acknowledge that there is still a long way to go with regard to accounting for the fine-grained structure in the data, especially with respect to recognition performance of the typical old items. For example, as reported in Table 4, the observed hit rates associated with the typical old items vary from as low as .58 to as high as .93, whereas the model's predictions for this class of items always hover between .80 and .82. It remains to be seen whether more sophisticated similarity-scaling approaches or alternative recognition models may improve on this level of prediction. We consider some possibilities along these lines in the General Discussion section.

As a source of comparison, we also fitted the traditional GCM to the data. The predictions of the main trends are shown along with the observed data in Table 5. In a nutshell, the traditional GCM fails to predict the old-item distinctiveness effect. Not surprisingly, the model yielded a fit to the individual-face data (BIC = 376.8) that was far worse than that of the HS-GCM.

In further explorations of the model, we considered the role of the different discrete-feature weight parameters (α, β, and γ) in Equations 8 and 10. Although including each parameter always led to an improvement in the BIC fit, the self-similarity γ parameter played the most important role, whereas the "correction" parameter β played a relatively minor role. At present, one of the major limitations of the modeling is that we do not have a good approach for representing similarity relations among the discrete features themselves. Thus, if two faces have distinctive eyes, we are unable to distinguish situations in which those eyes are themselves similar or dissimilar to one another. Future research is needed to develop more sophisticated modeling

approaches for taking into account this more fine-grained similarity structure.

Finally, we also considered the HS-GCM's predictions by using an alternative approach to combining the similarity and distinctiveness ratings of the faces. In this alternative approach, we fitted the similarity data themselves in terms of the hybrid-distance model given by Equation 8. By contrast, in the previous approach, the distinctiveness ratings were not included when fitting the MDS model to the similarity data. Thus, the "true" MDS solution could have been distorted because it was trying to capture the influence of the discrete-feature structure of the faces on the similarity judgments. The alternative approach, however, led to fits to both the similarity and the old–new recognition data that were virtually identical to the ones we have already reported. Thus, because none of our conclusions changed, we do not report the detailed results from this alternative mode of analysis.

GENERAL DISCUSSION

A fundamental challenge confronting global-familiarity models of recognition is the presence of old-item distinctiveness effects. Intuitively, because distinctive items lie in isolated regions of the psychological similarity space, their global familiarity would seem to be less than that of typical items that lie in dense regions. Thus, hit rates associated with distinctive items should be less than hit rates associated with typical items. However, the opposite result often seems to be observed, especially in the domain of face recognition (Busey & Tunnicliff, 1999; Valentine & Ferrara, 1991).

Old-item distinctiveness effects are far from universal, however, and seem to depend on the prevailing experimental conditions. For example, mere isolation in a continuous-dimension similarity space does not appear to be a sufficient condition for the emergence of the effects (see, e.g., Nosofsky & Zaki, 2003; Zaki & Nosofsky, 2001; see also Shiffrin, Huber, & Marinelli, 1995). Instead, Nosofsky and Zaki's (2003) research suggests that the presence of discrete individuating features is crucial.

To account for this pattern of results, Nosofsky and Zaki (2003) proposed a hybrid-similarity exemplar model of old–new recognition. In the hybrid model, similarity is a joint function of the distance between objects in a continuous-dimension space and of the extent to which the objects have matching and mismatching discrete features. The key to the modeling approach is that the feature-matching component makes allowance for differential self-similarities among objects. A distinctive object with highly salient discrete features has a greater degree of self-match than do typical objects without such features. Thus, the model accounts jointly for the findings that mere isolation in a continuous-dimension space does not give rise to robust old-item distinctiveness effects, whereas distinctiveness arising from the presence of discrete salient features does give rise to such effects.

Nosofsky and Zaki (2003) demonstrated preliminary support for the modeling ideas in a series of experiments in which distinctive alphanumeric characters were added to a set of continuous-dimension color patches. Besides accounting for the magnitude of the old-item distinctiveness effects across conditions, the model was also able to account for an intricate set of findings involving false alarm rates to foils that possessed distinctive features.

Nevertheless, this preliminary support for the model was obtained under exceedingly artificial conditions with highly impoverished stimuli. The central goal of the present research was to begin an exploration of the modeling ideas in the much more complex domain of face recognition. This domain seems especially appropriate in view of the fact that past findings from the face recognition literature have posed strong challenges to exemplar familiarity models.

In Experiment 1 of the present study, we used as stimuli a set of artificial faces to produce clusters with a high degree of within-cluster similarity and a low degree of between-cluster similarity. Furthermore, in analogy with the experiments conducted by Nosofsky and Zaki (2003), we manipulated whether or not old targets, high-similarity foils, and low-similarity foils contained discrete individuating features. Unlike Nosofsky and Zaki, we used types of discrete features that would plausibly occur on faces in the natural world (e.g., moles, scars, glasses, etc.). Despite moving to this much more complex domain, the discrete-feature manipulations led to much the same pattern of old–new recognition results as had been observed by Nosofsky and Zaki: A robust old-item distinctiveness effect was observed, and it became far easier to correctly reject low-similarity foils when they included a distinctive feature. In the present design, however, we found that high-similarity foils that included distinctive features had higher false alarm rates than did typical foils, a result that differed from those previously reported by Nosofsky and Zaki. However, the hybrid-similarity exemplar model accounted in quantitative detail for all of these effects. As we explained in detail earlier in this article, according to the hybrid model, adding discrete features to objects results in competing influences on overall summed similarity and global familiarity. The final outcome of the competition depends on various prevailing experimental conditions, including parametric variations in similarity among targets and foils, as well as the salience of the individuating discrete features themselves. In our view, it is a credit to the model that it can account for these intricate patterns of findings involving the effects of discrete distinctive features on false alarm rates.

In the present Experiment 2, we began explorations of the model in a still more challenging domain—namely, the old–new recognition of a set of naturalistic faces. In this domain, we no longer had strong control over within- and between-cluster similarity among faces, nor over the presence of discrete individuating features. Instead, we relied on observers' ratings of similarity and distinctiveness to provide approximate information along these lines. We then combined the ratings with the hybrid-similarity exemplar model to generate predictions of old–new recognition performance. The model yielded a fair quantita-

tive account of the probability with which the participants judged each *individual* face to be old or new. Furthermore, it accounted for the main qualitative finding that those faces judged to have the most salient discrete distinctive features had both the highest hit rates and the lowest false alarm rates.

In previous work, Dailey, Cottrell, and Busey (1998, 1999) also conducted modeling explorations of the naturalistic face recognition data from Busey and Tunnicliff (1999). They explored a variety of different combinations of modeling and similarity representation assumptions. The most successful was a combination of a kernel density estimation model with an MDS similarity representation. Although a detailed discussion goes beyond the scope of this article, the basic idea in the kernel density estimation model is that an aggregate "probability density" surface is built up over the multidimensional face space by summing the probability densities associated with each individual studied face. As noted by Dailey et al. (1999), the baseline version of this model is essentially the same as the GCM, with probability densities taking the place of individual-exemplar similarity gradients (see Ashby & Alfonso-Reese, 1995, for a discussion showing formal identities between kernel density estimation and summed-similarity exemplar models). However, Dailey et al. (1999) made special provision for old-item distinctiveness effects in the recognition data by assuming that items in isolated regions of the MDS solution had larger kernels. Thus, the aggregated probability density for distinctive faces could exceed that of more typical ones. In our view, an important limitation of this modeling approach is that it does not account for the findings of Zaki and Nosofsky (2001) and Nosofsky and Zaki (2003) that mere isolation in a continuous-dimension similarity space does not appear to yield robust old-item distinctiveness effects. Instead, the presence of salient discrete individuating features seems to be crucial. The hybrid-similarity exemplar model provides a natural account of such effects in terms of the discrete-feature matching mechanism providing boosts to self-similarity.

Clearly, future research is needed to improve upon the applications of the hybrid-similarity exemplar model in the domain of naturalistic face recognition. Although the model captured the broad qualitative pattern of results involving the different types of faces (distinctive, isolated, and typical), and provided a good account of the false alarm rates associated with the new faces, it failed to account for the hit rate variability associated with the subset of typical old faces. There are several likely reasons for this limitation. First, in the naturalistic face domain, there are likely to be various factors that influence recognition beyond the variables considered here. For example, a particular face might remind observers of some famous real-world face, which would exert a powerful influence on recognition judgments. Or, different emotional expressions might capture different degrees of attention, resulting in differential encodings of the individual faces. The present modeling approach would need to be extended to account for the effects of these complex variables.

Second, although it was a reasonable method for getting started, our view is that the use of direct individual-feature distinctiveness ratings for identifying discrete-feature structure is inadequate. For example, the underlying dimensional structure of a set of objects is almost certainly not fully accessible to consciousness, and people's verbal reports can provide at best an incomplete approximation of this structure. Likewise, as discussed earlier in our article, the technique is inadequate for representing the fine-grained similarity structure among the distinctive features themselves.

A preferable method, we believe, would involve the application of similarity-scaling techniques for extracting such structure. A highly promising direction might involve the application of a newly developed hybrid-similarity scaling algorithm proposed by Navarro and Lee (2003). The aim of this algorithm is to jointly extract continuous-dimension and discrete-feature structure through the analysis of matrices of similarity data. A good deal of work, however, is still needed for testing and fine-tuning the algorithm. With continued development, it might be possible to derive a hybrid-similarity scaling solution for a set of naturalistic faces and use this solution in combination with the exemplar model for predicting face recognition.

In sum, the central finding from this research is that the presence of discrete distinctive features on faces gives rise to robust old-item distinctiveness effects, such that hit rates associated with the distinctive faces greatly exceed hit rates associated with more typical faces. This result poses a challenge to standard global-familiarity models of recognition, because the global activation or similarity produced by a typical face should presumably be greater than that produced by a distinctive one. The solution proposed within the framework of the hybrid-similarity GCM, however, is that the self-match or self-similarity of a distinctive face is far greater than that associated with typical ones. If this self-match is sufficiently large, the global activation produced by a distinctive face can exceed that produced by typical faces. We obtained support for this interpretation in terms of formal quantitative fits of the model, both in a highly controlled experiment that involved manipulation of discrete-feature structure on a set of artificial faces and in an exploratory investigation involving a more complex set of naturalistic faces. Future research is needed to provide more rigorous tests of the modeling approach in the naturalistic face domain.

REFERENCES

ASHBY, F. G., & ALFONSO-REESE, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, **39**, 216-233.

BARTLETT, J. C., HURRY, S., & THORLEY, W. (1984). Typicality and familiarity of faces. *Memory & Cognition*, **12**, 219-228.

BUSEY, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, **9**, 476-483.

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.

DAILEY, M. N., COTTRELL, G. W., & BUSEY, T. A. (1998). Eigenfaces for

familiarity. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 273-278). Mahwah, NJ: Erlbaum.

DAILEY, M. N., COTTRELL, G. W., & BUSEY, T. A. (1999). Facial memory is kernel density estimation (almost). In M. S. Kearns, S. A. Solla, & D. A. Cohn, (Eds.), *Advances in neural information processing systems 11* (pp. 24-30). Cambridge, MA: MIT Press.

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*, **93**, 411-428.

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

KAYSER, A. (1985). *Heads*. New York: Abbeville Press.

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.

LEE, M. D., & NAVARRO, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, **9**, 43-58.

LIGHT, L. L., KAYRA-STUART, F., & HOLLANDER, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 212-228.

MALMBERG, K. J., & NELSON, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition*, **31**, 35-43.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

NAVARRO, D. J., & LEE, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67-74). Cambridge, MA: MIT Press.

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: Learning, Memory, & Cognition*, **17**, 3-27.

NOSOFSKY, R. M., & ZAKI, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old–new recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1194-1209.

O'TOOLE, A. J., ABDI, H., DEFFENBACHER, K. A., & VALENTIN, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A*, **10**, 405-411.

PHILLIPS, P. J., MOON, H., RIZVI, S. A., & RAUSS, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **22**, 1090-1104.

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

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

TVERSKY, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

VALENTINE, T., & FERRARA, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology*, **82**, 87-102.

VERGUTS, T., AMEEL, E., & STORMS, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition*, **32**, 379-389.

VOKEY, J. R., & READ, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, **20**, 291-302.

WASSERMAN, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92-107.

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.

NOTES

1. The last term ($33 \cdot s_b \cdot D$) comes from two sources: ($5 \cdot s_b \cdot D$) measures the summed similarity of the high-similarity distinctive foil to the 5 distinctive old items from the other 5 marked clusters, and the remainder ($28 \cdot s_b \cdot D$) measures the summed similarity of the high-similarity distinctive foil to the 28 typical old items from the other 11 clusters (2 typical old items from each of the 5 marked clusters and 3 from each of the 6 unmarked clusters). In this modeling approach, the parameter D represents both the mismatch between the presence of a distinctive feature and its absence, as well as the mismatch between two different distinctive features.

2. Based on the results of the distinctiveness ratings obtained in this experiment, two faces were reassigned from the distinctive group to the isolated group, and two from the isolated group were reassigned to the distinctive group (one old item and one foil from each group).

3. Of course, the extent to which a particular feature operates as a discrete distinctive feature will depend on the context of other items that are presented in the experiment. Thus, if all faces have mustaches that vary continuously in color, thickness, and so forth, then the mere presence of a mustache would no longer constitute a discrete distinctive feature. In the present case, we judged the discrete distinctive features to be either "present" or "absent" and not to vary continuously across other items in the set.

4. By definition, from Equations 8 and 9,

$$
\begin{aligned}
s_H(i,j) &= \exp\left\{-\kappa\left[d(i,j) + \alpha \cdot \left(D_i + D_j\right) - \beta \cdot \delta(i,j) \cdot \left(D_i + D_j\right)\right]\right\} \\
&= \exp\left[-\kappa \cdot d(i,j)\right] \cdot \exp\left[-\kappa \cdot \alpha \cdot \left(D_i + D_j\right)\right] \\
&\quad \cdot \exp\left[\kappa \cdot \beta \cdot \delta(i,j) \cdot \left(D_i + D_j\right)\right] \\
&= s(i,j) \cdot D_{ij} \cdot C_{ij},
\end{aligned}
$$

which is in the same form as Equation 3. In the present case, $s(i,j) = \exp[-\kappa \cdot d(i,j)]$ is the continuous component, $D_{ij} = \exp[-\kappa \cdot \alpha \cdot (D_i + D_j)]$ is the distinctive-feature mismatch component, and $C_{ij} = \exp[\kappa \cdot \beta \cdot \delta(i,j) \cdot (D_i + D_j)]$ is the common-feature match component.