Memory & Cognition
2001, 29 (8), 1165-1175

Category variability, exemplar similarity, and perceptual classification

ANDREW L. COHEN, ROBERT M. NOSOFSKY, and SAFA R. ZAKI
Indiana University, Bloomington, Indiana

Experiments were conducted in which observers learned to classify simple perceptual stimuli into low-variability and high-variability categories. Similarities between objects were measured in independent psychological-scaling tasks. The results showed that observers classified transfer stimuli into the high-variability categories with greater probability than was predicted by a baseline version of an exemplar-similarity model. Qualitative evidence for the role of category variability on perceptual classification, which could not be explained in terms of the baseline exemplar-similarity model, was obtained as well. Possible accounts of the effects of category variability are considered in the General Discussion section.

According to exemplar models of perceptual classification, people represent categories by storing individual exemplars in memory and classify objects on the basis of their similarity to these stored exemplars (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). Exemplar models have been successful at predicting a wide variety of perceptual classification phenomena, including details of classification learning, patterns of generalization to new transfer stimuli, and the time course of classification decision making. In the present research, however, we pursued an avenue that may demonstrate a fundamental limitation of these models.

The key previous study that motivated the present work was the classic set of experiments reported by Rips (1989) on the role of category variability in classification judgment. An example of one of these experiments is as follows. Participants were asked to imagine a circular object with a 3-in. diameter. One group of participants was asked whether this object was more likely to belong to the category of quarters or the category of pizzas. A second group of participants was asked whether the object was more similar to the category of quarters or the category of pizzas. (It had previously been determined that the 3-inch object lay midway between the largest quarter and the smallest pizza that a participant could remember.) The categorization group judged the object to be more likely to belong to the category of pizzas, whereas the similarity group judged the object to be more similar to the category of quarters. Thus, Rips demonstrated a dissociation between similarity judgment and category judgment and therefore concluded that categorization cannot be reduced to similarity. Smith and Sloman (1994) observed a similar pattern of results under certain experimental conditions, although their work also pointed to cases in which the generality of the effect was limited.

A key point of Rips's (1989) study is that observers' classification judgments are strongly influenced by their knowledge of the variability associated with alternative categories. Although the 3-in. object is judged as more similar to the QUARTER category, observers know that quarters display essentially zero variability in their size, whereas pizzas display a good deal of variability on this dimension. Thus, the observers' knowledge of category variability exerted an influence on classification decisions that went above and beyond similarity judgment per se. Furthermore, because numerous models of classification are based on similarity, including exemplar models, these experiments on the role of category variability pose an interesting challenge to such models.

In the present work, we focused on the pattern of classification judgments observed in Rips's (1989) experiments. Although the similarity-judgment question that Rips asked his participants is an intriguing one, various researchers have expressed concerns about the meaningfulness of this question (Goldstone, 1994; Nosofsky & Johansen, 2000). In addition, models of perceptual classification were not designed to predict how observers make direct judgments of object-to-category similarity. As will be seen, however, the classification results in and of themselves are sufficient to severely challenge the predictions of exemplar-similarity models, so this branch of Rips's study is the one that we pursued in the present research.

The representative of the class of exemplar-similarity models that we focused on in this study is the generalized context model (GCM; Nosofsky, 1986). In the GCM, exemplars are represented as points in a psychological space, and similarity between exemplars is a decreasing function of distance between the exemplars in the space. According to

This work was supported by Grant PHS R01 MH48494-09 from the National Institute of Mental Health to R.M.N. and by a Jacob Javits Fellowship to A.L.C. Correspondence should be addressed to A. L. Cohen, Department of Psychology, 1101 E. 10th St., Indiana University, Bloomington, IN 47405-7007 (e-mail: alcohen@indiana.edu), or to R. M. Nosofsky, Department of Psychology, 1101 E. 10th St., Indiana University, Bloomington, IN 47405-7007 (e-mail: nosofsky@indiana.edu).

the model, an observer sums the similarity of an item to the exemplars of the alternative categories, and the classification decision is based on the relative magnitude of these summed similarities. Formally, according to the baseline version of the model, the probability that item i is classified into Category A is given by

$$P(A|i) = \frac{\sum_{a \in A} M_a s_{ia}}{\sum_{a \in A} M_a s_{ia} + \sum_{b \in B} M_b s_{ib}}, \quad (1)$$

where s_{ia} denotes the similarity between item i and exemplar a, and M_a denotes the relative frequency with which exemplar a is experienced as a member of Category A. The terms $\Sigma M_a s_{ia}$ and $\Sigma M_b s_{ib}$ denote the summed similarities of item i to the exemplars of Categories A and B, respectively.

The similarity between item i and exemplar a is computed as follows. First, the distance between item i and exemplar a is given by

$$d_{ia} = \sqrt{\sum_m w_m (x_{im} - x_{am})^2}, \quad (2)$$

where x_{im} denotes the value of item i on psychological dimension m, and w_m is a weight parameter that reflects the degree of attention that an observer gives to dimension m in judging psychological distance. The similarity between item i and exemplar a is an exponential decay function of psychological distance (Shepard, 1987),

$$s_{ia} = \exp(-cd_{ia}), \quad (3)$$

where c is an overall scaling parameter.

To see why Rips's (1989) classification problems pose a fundamental challenge to the GCM, consider the abstract structure of the problems illustrated schematically in Figure 1. There are two categories of objects (A and B) that vary along a single dimension. Category A has very low variability along that dimension (it is analogous to the QUARTER category), whereas Category B has high variability (analogous to the PIZZA category). Item i is located midway between the exemplar with greatest magnitude from Category A and the exemplar with smallest magnitude from Category B. As is clear from Figure 1, item i is highly similar to all of the exemplars from Category A but is highly similar to relatively few exemplars from Category B. Thus, the summed similarity of item i to Category A exceeds its summed similarity to Category B, so the GCM predicts that observers should tend to classify such an object into the low-variability category. By contrast, the fundamental result observed by Rips was that participants tended to classify such objects into the high-variability category.[1]

Although Rips's (1989) results clearly challenge the predictions from the GCM, several aspects of his experiments make it difficult to draw strong conclusions with respect to the model. First, the GCM was designed primarily as a model of how participants learn categories of perceptual stimuli by induction over training exemplars. For example, the model has been used to predict the cate-

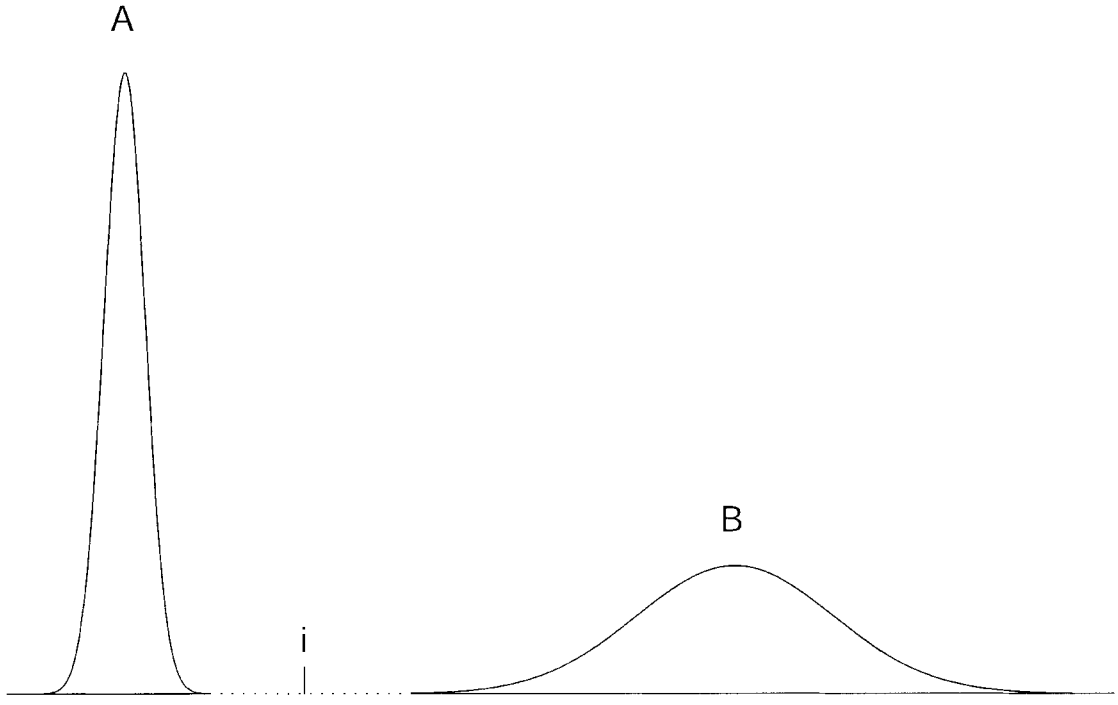

Figure 1. A schematic illustration of the type of category structure used by Rips (1989). Exemplar *i* is midway between the nearest exemplar of both categories.

gorization performance of observers viewing colors (Nosofsky, 1987), semicircles with radial lines (Nosofsky, 1986), dot patterns (Shin & Nosofsky, 1992), and simple geometric shapes (Nosofsky, 1984). Rips, however, asked participants to make judgments about categories of relatively high-level stimuli that were learned preexperimentally, such as teapots, automobiles, the number of members in the U.S. Senate, and the duration of a dinner party. It remains an open question how situations involving prior knowledge and highly conceptual types of categories should be modeled within the framework of the GCM and other similarity-based models (for some ideas along these lines, see Goldstone & Barsalou, 1998, and Heit, 1994).

Furthermore, as noted by Rips (1989), "although the mystery item was numerically midway between the subjects' extreme values, it may have been subjectively closer to the larger category" (p. 31). This possibility does not in itself explain the dissociation between similarity and categorization judgments reported by Rips. However, as we discussed earlier, we are reluctant to rely on the results of observers' direct ratings of the similarity of an object to a category to assess the model's predictions. Because the GCM bases its categorization predictions on the psychological similarities among individual exemplars, it is critical that this psychological similarity space be carefully mapped out.

The general goal of our experiments, therefore, was to test paradigms that conceptually replicated the structure of Rips's (1989) classification problems, but in situations within the GCM's intended domain of inquiry. Thus, in the present experiments, rather than considering prelearned highly conceptual categories, participants learned simple perceptual categories by induction over individual training exemplars. In addition, to more rigorously assess the predictions from the GCM, we conducted similarity-scaling experiments to precisely determine the locations of the individual exemplars in psychological space. The central question was whether or not there would be effects of category variability on perceptual classification performance, analogous to the results observed by Rips, that violated the predictions from the exemplar-based GCM.

PRELIMINARY EXPERIMENTS

In a series of preliminary experiments, we sought to replicate the classification results from Rips's (1989) QUARTER–PIZZA experiments, except in simple perceptual domains in which participants learned the categories by induction over training exemplars. The design of one such experiment is illustrated schematically in Figure 2. The stimuli were vertical lines varying only in length. A low-variability category, analogous to QUARTER, was defined by Line 1 in the figure; whereas a high-variability category, analogous to PIZZA, was defined by Lines 2–8. In a classification training phase, the participants learned to associate the appropriate category label with each individual line. To equate the category frequencies, Line 1 was shown seven times as often as were each of Lines 2–8 during training. Following training, the participants were tested in a transfer phase that included all of the old exemplars and that included presentations of the critical "middle" stimulus illustrated in Figure 2 by the line labeled "MS." (We had conducted extensive similarity-scaling experiments to find a middle stimulus that was equally similar to the nearest exemplars of the low- and high-variability category; for details, see the Procedure section of Experiment 1.) The result we obtained was that the participants were more likely to classify the middle stimulus into the low-variability category than into the high-variability category. Thus, we failed to provide a demonstration analogous to the classification results observed in Rips's experiments.

In other preliminary experiments, we tried to control for possible "edge" effects involving these unidimensional stimuli (Braida & Durlach, 1972), and for any directional biases that might be present in observers' judgments. For example, in one experiment, the low-variability category was defined by a single line length located in the center of the stimulus range, whereas the high-variability category was defined by distributions of line lengths located on both sides of this center line. Extensive similarity-scaling work was performed to find middle stimuli that were equally similar to the line from the low-variability category and to

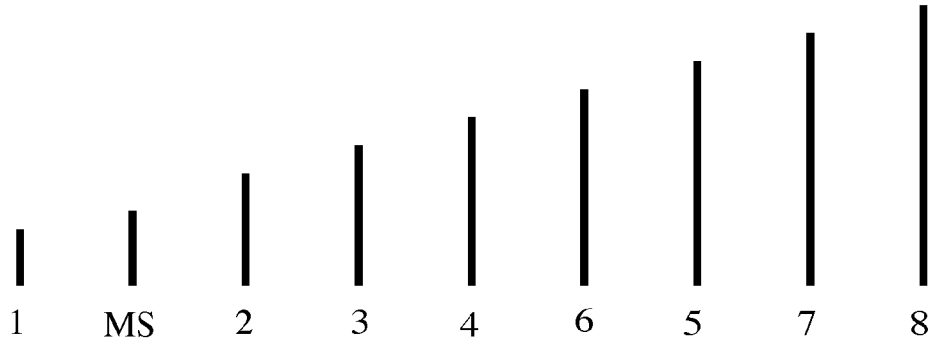

Figure 2. A schematic illustration of one of the preliminary experiments designed to conceptually replicate Rips (1989). Line 1 was the low-variability category, and Lines 2–8 were the high-variability category. The middle stimulus, MS, was scaled to be perceptually midway between Lines 1 and 2.

the nearest neighbors from the high-variability category. Again, however, the participants tended to classify these middle stimuli into the low-variability category rather than into the high-variability category, in contrast to the phenomenon observed by Rips (1989).

We were limited in the conclusions that we could reach based on the qualitative pattern of results from these preliminary experiments. Following the schematic design of Rips's (1989) experiments, these preliminary experiments basically pitted two variables against one another. Although the middle stimulus had equal psychological spacing to the closest exemplars of the competing categories, its summed similarity was greater to the low-variability category than to the high-variability one. Even if variability per se had indeed exerted an effect on categorization judgments, the designs may simply not have revealed it because the competing variable of summed similarity exerted too powerful an opposite effect.

Thus, rather than continue to search for designs that might replicate the qualitative effects demonstrated by Rips (1989), we pursued some alternative research strategies. The first strategy, pursued in Experiment 1, was to investigate potential effects of category variability on classification performance by manipulating it across conditions as an independent variable, while trying to hold constant across conditions the competing variable of summed similarity. The second strategy, pursued in Experiment 2, was to conduct detailed quantitative tests of whether or not the GCM could predict classification choice probabilities in situations involving categories of differing variability.

EXPERIMENT 1

The purpose of Experiment 1 was to manipulate category variability across conditions as an independent variable, while trying to hold constant across conditions the competing variable of summed similarity. Unfortunately, we know of no method that allows one to perfectly satisfy this goal. Any manipulation of category variability across conditions will by necessity change the exemplar similarity relations that are involved (as defined in the GCM). However, it is possible to achieve a close approximation to the desired manipulation.

Recall that, according to the GCM, the similarity between two items is an exponentially decreasing function of the distance between those items in multidimensional space (Equation 3). One property of the exponential is that it decays very rapidly, so that items far from the middle stimulus will add very little to the relevant summed similarity calculations. The key is to add high-variability category exemplars that are sufficiently far from the middle stimulus so as to increase the category variability without much influencing the summed similarity.

We used a category structure very similar to that of our first pilot experiment. There were two conditions. In both conditions, both the low-variability category exemplar and the middle stimulus were identical to those from the first pilot study (see Figure 2); however, the nature of the high-variability category changed across conditions. In Condition 1, the high-variability category consisted of only the two smallest exemplars of the high-variability category from the pilot experiment (i.e., Line 2 and Line 3 in Figure 2). In Condition 2, the high-variability category contained all seven high-variability category exemplars (i.e., Lines 2–8). The main difference between Conditions 1 and 2 is the addition of five stimuli to the high-variability category that are relatively distant from the middle stimulus. As will be seen, these additional stimuli increase the variability of the high-variability category, while leaving the summed similarity of the middle stimulus to the high-variability category virtually unchanged. Thus, the GCM predicts almost identical classification probabilities for the middle stimulus across the conditions. By contrast, if there is an effect of variability on categorization analogous to the phenomenon reported by Rips (1989), then the middle stimulus should be classified into the high-variability category more often in Condition 2 than in Condition 1.

Note that, in the present design, the middle stimulus was more centrally located with respect to the entire range of stimuli in Condition 1 than in Condition 2. Because of possible edge or range effects on unidimensional perception (Braida & Durlach, 1972; Luce, Nosofsky, Green, & Smith, 1982), we were concerned that the middle stimulus might be psychologically more similar to the nearest exemplar from the high-variability category in Condition 2 than in Condition 1.[2] Such a change in similarity relations would also result in the prediction that the middle stimulus would be classified into the high-variability category more often in Condition 2. To determine the extent of any change in similarity relations across conditions, a separate group participated in an identification task involving the same stimuli and stimulus frequencies as in the categorization tasks. An appreciable change in the identification pattern of the middle stimulus across conditions would indicate a shift in the similarity structure of the stimulus space and make it difficult to separate out any effects of variability and similarity in the classification task. The absence of such an effect would support the idea that any change in the pattern of classification judgments across conditions was due to an effect of variability per se.

Method

Participants. Thirty-five Indiana University undergraduates participated in each condition of the identification task. Thirty-five and 37 Indiana University undergraduates participated in Conditions 1 and 2 of the categorization task, respectively. All received course credit for their participation.

Stimuli. The same stimuli were used in the corresponding conditions of the identification and classification tasks. The stimuli were lines of varying length. The bottom point of each line was vertically and horizontally centered on the screen. The full set of training stimuli consisted of lines of pixel length 30, 60, 75, 90, 105, 120, 135, and 150. Condition 1 of both tasks used only the first three of these training stimuli. Condition 2 of both tasks used all eight of these training stimuli. For ease of description, we will often refer to these line stimuli in terms of their pixel length.

The middle stimulus in both conditions of both tasks was a line of 39 pixels, the same middle stimulus used in our first pilot experiment (see Figure 2). Stimulus 39 was chosen as the middle stimulus

for two reasons. First, extensive similarity-scaling work that we conducted during our pilot experiments had indicated that this line was near the center of the subjective range between Stimuli 30 and 60, thereby lessening the chance of any ceiling or floor effects. Second, pilot work showed that, in tasks of unidimensional identification involving these lines, Stimulus 39 was rarely confused with lines of length greater than 60 (less than 2.0% of the trials). This result adds weight to the argument that the additional 5 stimuli in Condition 2 should leave the summed similarity calculation from the exemplar model relatively unchanged.

Procedure. Both the identification task and the classification task were organized into training and transfer phases. Both tasks followed the same general structure. There were 150 trials in the training phase. Stimuli 30 and 60 were both shown on 45% of the training trials, and the remaining 10% of the trials were split equally among the remaining training stimuli. Thus, the bulk of training was on the two stimuli nearest to the middle stimulus, so the addition of the training stimuli far from the middle stimulus would change the summed similarity calculation even less. In the identification task, the participants learned a unique label for each training stimulus via feedback. In the classification task, the participants learned via feedback that Stimulus 30 was in one category (the low-variability category), and the remaining training stimuli were in a second (high-variability) category.

After training, the participants began the transfer phase. This phase consisted of 200 trials. The middle stimulus was seen on 10% of the transfer trials, and the presentation rates for each of the training stimuli dropped by 10%. The participants received feedback after each training stimulus. Following responses to the middle stimulus, the participants received a message of "Thank You."

In both the identification and classification training and transfer phases, stimuli were selected for presentation randomly on each trial within the constraints stated above.

Results

In the identification task, the overall transfer phase error rate for the training stimuli was 6.1%. In Conditions 1 and 2, the middle stimulus was identified as Stimulus 60 or greater (which were the high-variability category exemplars in the classification task) on 43.0% and 42.9% of the transfer trials, respectively ($MS_e = 577.80$). A t test confirmed that the participants were not changing the identification pattern of the middle stimulus across conditions [$t(68) = 0.031, p = .975$], thus indicating that psychological similarity relations involving the middle stimulus were essentially unchanged across the variability conditions.

In the categorization task, the transfer phase error rates for Stimuli 30 and 60 averaged across all participants and both conditions were 4.5% and 2.0%, respectively. One participant from Condition 1 had a transfer phase percent correct for the training stimuli of less than 75% and was excluded from all further analyses. In Conditions 1 and 2, the middle stimulus was classified into the high-variability category on 29.5% and 47.1% of the trials, respectively ($MS_e = 773.62$). A t test confirmed that the middle stimulus was judged to be a member of the high-variability category significantly more often in Condition 2 than in Condition 1 [$t(70) = 2.693, p < .01$].

Discussion

In Experiment 1, we found qualitative evidence that category variability affects perceptual classification performance beyond what is predicted by the GCM. Although the summed similarity of the middle stimulus to the high-variability category exemplars remained essentially constant between conditions, we found that classification performance changed dramatically. Specifically, the middle stimulus was more likely to be classified into a category as the variability of that category increased. Furthermore, the results cannot be attributed to a changed similarity structure across the two conditions, because, in the corresponding identification tasks, there was no effect of the variability manipulation on the identifications of the middle stimulus.

In sum, these data suggest that the baseline version of the GCM—that is, a version in which classification judgments are based solely on relative summed similarity to stored exemplars—fails to account fully for the effect of category variability. The results echo the findings from Rips (1989) in that high category variability exerts a "pull" on classification judgments that goes beyond the influence of exemplar-based similarity alone.

EXPERIMENT 2

In Experiment 2, we used another approach to test for effects of category variability. First, in our pilot studies and in Experiment 1, the stimuli were always unidimensional line lengths. For purposes of generality, in Experiment 2, we instead used stimuli varying along two dimensions: colors varying in brightness and hue. Second, in Experiment 2, our research strategy was to test whether the GCM could quantitatively predict observers' choice probabilities in situations in which category variability was manipulated. We hypothesized that the GCM would systematically underpredict the probability with which a middle transfer stimulus was classified into high-variability categories.

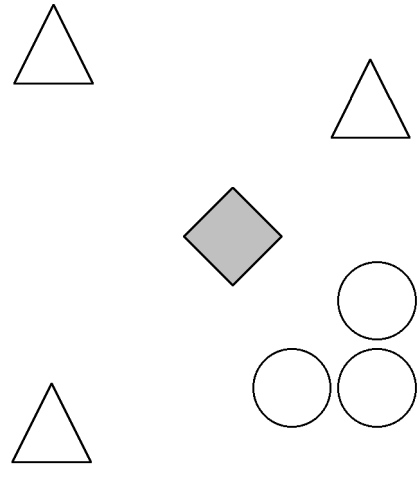

Figure 3. A schematic illustration of the color space for Experiment 2. The triangles are the high-variability category members, the circles are the low-variability category exemplars, and the diamond is the middle stimulus. The location of the low-variability category exemplars was rotated across conditions.

A schematic illustration of the type of condition tested in Experiment 2 is shown in Figure 3. During a training phase, the participants learned that three of the training exemplars (the triangles in the figure) were members of the high-variability category and that the other three training exemplars (the circles) were the low-variability category exemplars. Each of the training exemplars was presented with equal frequency, so the categories had equal base rates. In a subsequent transfer phase, the participants classified the training stimuli and the middle transfer stimulus (the diamond in Figure 3). Extensive pilot work was conducted to find configurations in which the middle transfer stimulus was roughly equally similar to the exemplars of the low-variability category and the nearest exemplar of the high-variability category.

Note from Figure 3 that the middle stimulus has greater summed similarity to the low-variability category than to the high-variability category. Thus, to the extent that the participants tended to classify the middle stimulus into the high-variability category, rather than the low-variability one, it would constitute qualitative evidence against a pure exemplar-similarity account of categorization. Beyond seeking this qualitative evidence, our main goal was to test the ability of the GCM to quantitatively account for the full set of transfer data. We derived multidimensional scaling (MDS) solutions for the colors in each condition and then used the GCM in combination with the scaling solutions to quantitatively fit the data.

Finally, to ensure that any tendency to classify the middle stimulus into the high-variability category was due to the variability manipulation per se, and not to some stimulus-specific properties associated with the high-variability exemplars, four different conditions were tested in which the relative locations of the categories in the color space were varied. In each condition, the cluster of low-variability exemplars was placed in a different one of the four corners of the color space illustrated in Figure 3.

Table 1
Red (R), Green (G), and Blue (B) Values for Each Stimulus in Experiment 2

Stimulus	R	G	B
1	222	225	255
2	212	225	255
3	197	230	255
4	190	234	253
5	195	196	235
6	159	201	227
7	152	176	215
8	145	142	189
9	110	153	186
10	115	114	168
11	96	118	168
12	62	122	168
13	38	122	161

Note—Stimulus 7 was the middle stimulus. Stimuli 1, 4, 10, and 13 were shown in all conditions. Stimuli 9 and 12 were shown in Condition 1 only. Stimuli 8 and 11 were presented in Condition 2 only. Stimuli 2 and 5 were displayed in Condition 3 only. Stimuli 3 and 6 were shown in Condition 4 only.

Method

Participants. Eighty-eight Indiana University undergraduates participated in this experiment for course credit. There were 22 participants in Condition 1, 22 in Condition 2, 21 in Condition 3, and 23 in Condition 4. All claimed to have normal color vision.

Stimuli. The stimuli were 13 computer-generated colors. The particular colors that were used varied across four conditions of testing. According to the Munsell system, the colors varied primarily in hue and brightness and were of roughly the same saturation. A complete list of the Adobe Photoshop red, green, and blue values is given in Table 1. The precise psychological configuration of the colors in each condition is presented in the Results section.

The colors ranged from light purple in one corner to dark blue in the opposite corner. Each color occupied a 5.08×5.08 cm square on a white background. In the classification task, a single color was presented centered on the screen. In a similarity-scaling task, two color squares were shown, separated by a horizontal distance of 4.1 cm.

Procedure. We tested four separate categorization conditions, with the location of the cluster of low-variability category exemplars rotated across conditions (for an illustration, see Figure 4). In a training phase, the participants learned the category structure by induction over exemplars. On each trial, a color was shown, and the participant judged its category assignment. Feedback was given after each training trial. There were 20 training blocks. During each block, each low- and high-variability category exemplar was seen once. Following training, there were 20 transfer blocks. Each transfer trial proceeded exactly as in training; however, one trial was added per block. On this additional trial, the middle stimulus was displayed, and the participants received feedback of "Thank You." The order of presentation of stimuli within both training and transfer blocks was randomized.

To verify that the arrangement of colors in psychological space corresponded reasonably well to the schematic design illustrated in Figure 3, we conducted similarity-scaling studies following the completion of the classification task. These scaling studies were used to derive MDS solutions for the colors. The participants in Conditions 1 and 3 rated the similarities of all of the colors from Conditions 1 and 3. Likewise, the participants in Conditions 2 and 4 rated the similarities of all colors from Conditions 2 and 4. (MDS solutions were derived separately for these pairs of conditions, because we had originally planned to run only Conditions 1 and 3.) In each scaling study, there were four blocks of trials, and each possible pair of different stimuli was seen once per block. The order of the pairs and left–right placement of the colors on the screen were randomized. On each trial, the participants rated the similarity of the two colors on a 9-point scale (1 = *highly dissimilar*, 9 = *highly similar*). A rating scale was shown on the bottom of the screen as a reminder, and the participants were urged to use the entire scale range.

Results

Similarity scaling. The standard Euclidean model from ALSCAL was used to derive MDS solutions from the averaged similarity ratings. Averaged across conditions, the two-dimensional scaling solutions yielded a stress equal to .022 and accounted for 99.7% of the variance in the averaged similarity ratings. The resulting locations of the colors in two-dimensional psychological space are illustrated separately for Conditions 1–4 in Figure 4. In Conditions 2, 3, and 4, the configurations satisfied our goal of creating multidimensional category structures in which the middle stimulus was roughly equally similar to the cluster of exemplars of the low-variability category and the nearest exemplar of the high-variability category. This goal was not satisfied, however, in Condition 1. Although we

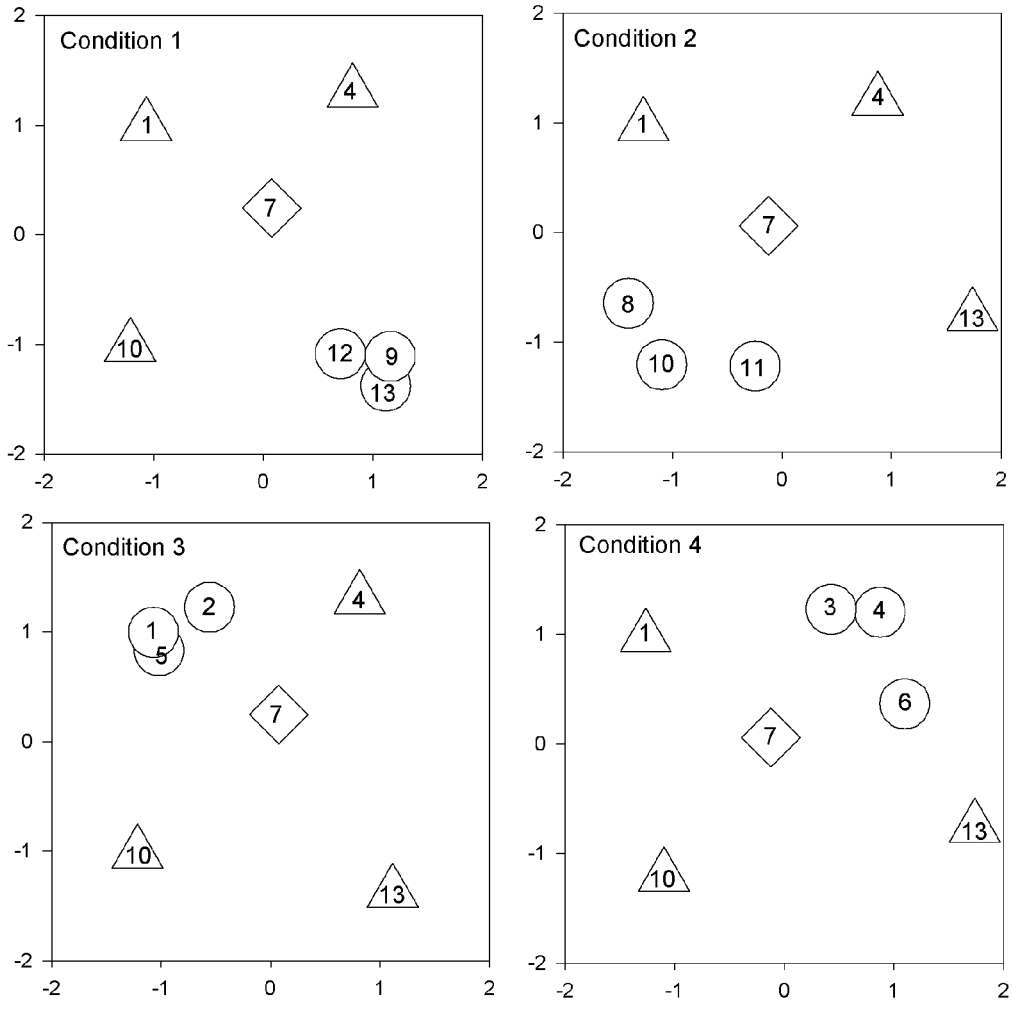

Figure 4. Multidimensional scaling (MDS) solutions derived from averaged similarity ratings for the stimuli of Experiment 2. The center of each shape represents the MDS coordinate for each color. For each condition, the triangles are the high-variability category exemplars, the circles are the low-variability category exemplars, and the diamond is the middle stimulus.

will not be able to evaluate the Condition 1 results with regard to the qualitative direction in which the middle stimulus was classified (because both the exemplar-similarity and variability factors favor the high-variability category), the data from this condition are still valuable for evaluating the quantitative predictions from the exemplar model.

Classification. The average error rates for training stimuli during transfer for Conditions 1, 2, 3, and 4 were 6.1%, 4.4%, 3.6%, and 5.0%, respectively. Two participants from Condition 3 had a transfer error rate on the training stimuli greater than 25% and were excluded from further analysis.

The middle stimulus was classified into the high-variability category in Conditions 1, 2, 3, and 4 on 73.0%, 50.7%, 55.8%, and 73.4% of the transfer trials, respectively. In Conditions 1 and 4, the middle stimulus was classified into the high-variability category on significantly greater than 50% of the transfer trials [$t(21) = 3.769, p < .01$, and $t(22) = 3.819, p < .01$, respectively]. In Conditions 2 and 3, the percentage of trials in which the middle stimulus was judged to be a member of the high-variability category did not differ significantly from 50% [$t(21) = 0.098, p = .923$, and $t(18) = 0.687, p = .501$, respectively].

Modeling results. The GCM was fitted to the classification transfer data from Conditions 1–4 by using a maximum-likelihood criterion. The coordinate values used in the distance function (Equation 2) were those that were derived from the MDS analysis of the similarity-ratings data. The modeling analyses assumed that the participants weighted equally the two psychological dimensions. Thus, the only free parameter in the model was the overall sensitivity parameter, c.

The modeling results are summarized in Table 2 in terms of the observed and predicted probability with which the middle stimulus was classified into the high-variability

category. In all four conditions, the baseline version of the GCM greatly underpredicted the observed classification probability of the middle stimulus. The departures of the observed data from the quantitative predictions of the model were statistically significant in Conditions 1, 3, and 4 [Condition 1, $t(21) = 2.456, p < .05$; Condition 2, $t(21) = 1.245$, $p = .227$; Condition 3, $t(18) = 2.349, p < .05$; Condition 4, $t(22) = 6.259, p < .01$].

Table 2
Average Proportion of Transfer Trials in Which the Middle Stimulus Was Classified Into the High-Variability Category in Experiment 2

Condition	Data	GCM
1	.73	.58
2	.51	.42
3	.56	.36
4	.73	.35

Discussion

In summary, in all four conditions of Experiment 2, the baseline version of the GCM underestimated the probability with which the middle test stimulus was classified by observers into the high-variability category. The departures of the observed classification proportions from the predictions of the model were statistically significant in Conditions 1, 3, and 4. The results were particularly dramatic in Condition 4. Here, the GCM predicted that observers would tend to classify the middle stimulus into the low-variability category, yet they classified it into the high-variability with probability significantly greater than .50. This result constitutes strong qualitative evidence against the predictions from the model.

One might try to save the exemplar-similarity model by positing that the observers may have differentially weighted the psychological dimensions composing the exemplars, whereas our modeling analyses assumed equal weighting of the two psychological dimensions. In the present case, we believe that making allowance for differential weighting of the psychological dimensions provides an ad hoc account of the results. A central assumption in past applications of the GCM is that, following extensive training experience with a category, observers learn to distribute attention across the psychological dimensions in a manner that tends to optimize performance (i.e., to maximize their percentage of correct classifications of the training stimuli). Although a detailed discussion goes beyond the scope of the present article, it turns out that the attention weights that are needed to allow the GCM to fit the classification transfer data are often highly suboptimal. For example, in some conditions, fitting the classification probability for the middle stimulus required the model to place most of its attention weight on the dimension that was less diagnostic of category membership. Because the underpredictions of the GCM were so systematic across the four conditions of testing, a much more plausible account of the results is that the variability manipulation exerted an effect on classification judgments in a manner that goes beyond the predictions from the baseline version of the model.

GENERAL DISCUSSION

Summary

In summary, our experiments provide evidence that there are systematic effects of category variability on perceptual classification that go beyond the predictions from the baseline version of the exemplar-similarity model. In Experiment 1, we arranged a design in which the variability of a target category increased across conditions, whereas the summed similarity of the middle stimulus to the category exemplars was held essentially constant across these conditions. Contrary to the predictions from the baseline GCM, the probability with which the observers classified the middle stimulus into the target category increased as the category's variability was increased. In Experiment 2, the baseline GCM systematically underpredicted the probability with which a transfer "middle" stimulus was classified into high-variability categories. In one case, the results were particularly dramatic: Because the middle stimulus had greater summed similarity to the low-variability category exemplars than to the high-variability ones, the GCM predicted that the middle stimulus would tend to be classified into the low-variability category rather than into the high-variability category. By contrast, the results went significantly in the opposite direction.

In this article, for purposes of brevity, we focused on the GCM's underprediction only in the four conditions of Experiment 2. We should emphasize, however, that this underprediction was an extremely robust finding that held in numerous other analogous experimental conditions that we did not report in detail. For example, following methods described by Nosofsky (1985), we derived unidimensional scaling solutions for the line-length stimuli used in our preliminary experiments and in Experiment 1. (These scaling solutions were derived by fitting a model known as the *MDS-choice model* to the identification confusion data collected in these experiments.) We then used the GCM in combination with the derived unidimensional scaling solutions to predict the probability with which the middle stimulus was classified into the high-variability category. In addition, we tested two-dimensional conditions analogous to those in our Experiment 2, except where the low-variability category consisted of a single high-frequency exemplar instead of the cluster of three exemplars illustrated in Figure 3. In every condition of every experiment that we conducted (a total of 13 conditions across five experiments), the GCM underpredicted the probability with which the middle stimulus was classified into the high-variability category.

Our results are consonant with the earlier findings of Rips (1989), whose influential study suggested strongly that there are effects of category variability on classification judgments that go beyond the predictions from pure "similarity-based" models. Our experiments, however, may

be viewed as providing even more direct evidence against the pure similarity-based version of the GCM than did Rips's results. First, Rips's findings were obtained in a situation in which people made classification judgments about prelearned and sometimes abstract conceptual categories. By contrast, the GCM was formalized with low-level perceptual stimuli in mind and for situations in which learning takes place via induction over individually presented category exemplars. Such conditions were strictly maintained in the context of the present experiments. Thus, the present challenges to the baseline GCM occurred in the type of domain for which the model was actually formalized.

The present demonstrations go beyond the ones reported by Rips (1989) in another important manner as well. To obtain evidence against similarity-based models of classification, Rips relied on the results from a task in which observers judged the similarity of a hypothetical item to entire categories of objects. Although the category-similarity judgment task is intriguing, the meaningfulness and interpretation of the task has nevertheless been questioned by investigators such as Goldstone (1994), Nosofsky and Johansen (2000), and others. By contrast, in the present experiments, similarities between exemplars were measured by using the same fundamental, psychological scaling techniques as in most past applications of the GCM. Thus, in the present experiments, the failures of the baseline model occurred in a situation in which we played by the model's own ground rules.

Accounting for the Effects of Category Variability

We conclude our article by considering possible accounts of the effects of category variability, although future research will be needed to distinguish among these possibilities.

First, as suggested by Smith and Sloman (1994), observers' tendency to classify transfer stimuli into high-variability categories rather than into low-variability ones may indicate that multiple systems underlie category judgment. Although similarity comparisons to stored exemplars may be one important component of categorization, observers may also make use of explicit rules. Intuitively, an observer may form a rule that an object must be virtually identical to the members of a low-variability category in order to be classified into that category. A variety of modern models of classification posit that categorization is governed by multiple systems, such as rules and exemplars (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994) or rules and procedural learning systems (Ashby, Alfonso Reese, Turken, & Waldron, 1998). It would be interesting to test whether these models predicted the category-variability effects observed in the present experiments.

Another possible explanation of our results springs from the class of "decision-boundary" models of Ashby and Maddox and their colleagues (Ashby & Lee, 1991; Ashby & Maddox, 1993; Maddox & Ashby, 1993). According to certain of these models, observers assume that categories are multivariate normally distributed, and, during training, they estimate the means, variances, and covariances of the category distributions. Observers then construct decision boundaries to optimally separate perceptual space into response regions. According to such models, the middle transfer stimulus tested in our designs would have a greater likelihood of having been generated from the high-variability category distribution, so the optimal decision boundary would classify it into that category. Indeed, in our Experiment 2 conditions, the middle stimulus would be closer to the centroid of the high-variability distributions, so it would have a much higher likelihood of having been generated from those categories. Perhaps observers do not always precisely estimate the parameters of the category distributions, which would explain why the middle stimulus was not always classified into the high-variability category in our experiments.

Finally, it may be possible to explain the category-variability effects while staying within the framework of exemplar-similarity models. Throughout our article, we have focused on the baseline version of the GCM formalized in Equation 1. A fuller version of the model, however, makes provision for the role of category response-bias parameters. As it was originally formalized (Nosofsky, 1986, 1987), according to the GCM, the probability that item i is classified into Category A is given by

$$P(A \mid i) = \frac{b_A \sum_{a \in A} s_{ia}}{b_A \sum_{a \in A} s_{ia} + (1 - b_A) \sum_{b \in B} s_{ib}}, \qquad (4)$$

where b_A ($0 \leq b_A \leq 1$) denotes the response-bias associated with Category A. Thus, a natural question is whether participants' tendency to classify objects into the high-variability category may reflect some form of response bias.

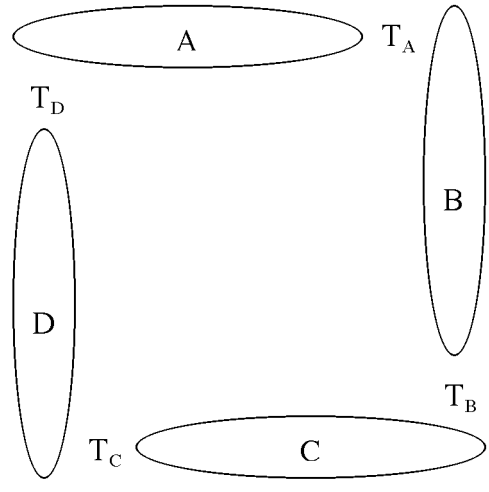

Figure 5. Schematic illustration of an experimental design to test whether observers are sensitive to the direction of variability. See text for details.

Admittedly, explanations based on response bias are often not very satisfying. In the present case, for example, it is obvious that, by assuming that there is a response bias associated with the high-variability category, the GCM will be able to better predict the probability with which the middle stimulus was classified into that category. Unless there is a principled reason, however, for positing why such a bias may exist, the explanation is post hoc and serves only to redescribe the data.

The reason that a response-bias explanation provides a potentially interesting account of our results, however, is that, in general, it is optimal for an observer to have a higher bias for responding with a high-variability category than with a low-variability one. In particular, such a pattern of response bias will often maximize the percentage of correct classification decisions that are made by the observer. The construct of parameter optimization has always been a central idea in theorizing involving the GCM. For example, Nosofsky (1984, 1986, 1987, 1991) provided evidence that observers will often distribute attention across the psychological dimensions that compose the exemplars in a manner that tends to optimize performance. In a similar vein, it seems reasonable to posit that observers may adjust their response biases to achieve such a goal. Indeed, the hypothesis that observers may have a tendency to adjust their response biases so as to optimize payoffs in alternative task settings has always been one of the central ideas underlying signal-detection analyses of classification data as well (Green & Swets, 1966).

Why is it generally optimal for an observer to have a response bias toward the high-variability category? To answer this question intuitively, consider the category structure illustrated in Figure 1. Consider in particular the stimulus from the high-variability category that lies closest to the low-variability one. This stimulus has relatively few neighbors from its own category to which it is highly similar, yet it is fairly similar to numerous exemplars from the low-variability category. To counteract this strong competition from the low-variability category exemplars, an observer would need to establish a response bias toward the high-variability category in order to classify this stimulus with high accuracy. Although such a response bias would also result in lowered accuracy for the members of the low-variability category, the adverse effect would be minimal. The reason is that each member of the low-variability category has numerous close neighbors from its own category that provide it with similarity-based support. The resulting high summed similarity for the low-variability category members counteracts the response bias toward the high-variability category. Thus, by adopting a response bias toward the high-variability category, an observer tends to optimize performance. We have verified that, for all of the category structures tested in this study involving a high-variability category, it is indeed optimal for an observer to have a higher response bias toward the high-variability category than toward the low-variability one.

A response-bias explanation may also mesh nicely with the original phenomena reported by Rips (1989). As noted earlier, observers judged a 3-in. disk as being more similar to the QUARTER category but were more likely to classify the object as a member of the PIZZA category. If one accepts the idea that the observers had a strong response bias to classify objects into the high-variability PIZZA category, then these results seem very sensible. The operation of a response bias should not be expected to influence observers' judgments of similarity, so the finding that the participants judged the 3-in. object as more similar to the QUARTER category than to the PIZZA category seems quite reasonable.

Thus, a critical goal for future research is to test whether there are effects of category variability on perceptual classification that go beyond the ability of an exemplar-similarity model to explain, even if provision is made for the possibility of differential response bias. To sketch one idea that we have along these lines, consider the design illustrated in Figure 5. There are four category distributions: A, B, C, and D. Distributions A and C exhibit high variability along the horizontal dimension, whereas Distributions B and D exhibit high variability along the vertical dimension. Four transfer stimuli are located midway between the nearest exemplars of adjacent category distributions. To the extent that observers are sensitive to the direction of variability in each distribution, we hypothesize that they would tend to classify T_A into Category A, T_B into Category B, and so forth. Even if allowance were made for the operation of category response biases, the standard exemplar-similarity model would be unable to account for such a pattern of results. To predict that T_A is classified into Category A rather than Category B, the response bias for A (b_A) would need to be set greater than the response bias for B (b_B), $b_A > b_B$. Likewise, the model would require $b_B > b_C$, $b_C > b_D$, and $b_D > b_A$. But this pattern violates the law of transitivity, and so it is impossible to satisfy this set of response-bias requirements. It is an open question whether the hypothesized pattern of classification transfer would be observed in such a design, but to the extent that it was, it would present a severe challenge even to versions of the exemplar-similarity model that make provision for category response bias.

REFERENCES

Ashby, F. G., Alfonso Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, **105**, 442-481.

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, **120**, 150-172.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, **51**, 483-502.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, **127**, 107-140.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

GOLDSTONE, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, **52**, 125-157.
GOLDSTONE, R. L., & BARSALOU, L. W. (1998). Reuniting perception and conception. *Cognition*, **65**, 231-262.
GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
HEIT, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1264-1282.
HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.
LUCE, R. D., NOSOFSKY, R. M., GREEN, D. M., & SMITH, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics*, **32**, 397-408.
MADDOX, W. T., & ASHBY, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, **53**, 49-70.
MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
NOSOFSKY, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.
NOSOFSKY, R. [M.] (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, **38**, 415-432.
NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
NOSOFSKY, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 87-108.
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., & JOHANSEN, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, **7**, 375-402.
NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.
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. (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.

NOTES

1. Note that a related study conducted by Fried and Holyoak (1984) does not pose the same direct challenge to the GCM as does Rips's (1989) study. In Fried and Holyoak's experiment, participants learned to discriminate between the members of a low-variability category and a high-variability one. At time of transfer, participants were tested on objects that were closer to the centroid of the low-variability category but that were statistically more likely to have been generated from the high-variability category. Because participants tended to classify these objects into the high-variability category, Fried and Holyoak's results provided evidence that people are sensitive to the distributional character of categories. However, this general pattern of results is consistent with the predictions from exemplar models. Given that the critical transfer stimuli were more likely to have been generated from the high-variability category, it is also likely that they were more similar to specific training exemplars of the high-variability category than to those of the low-variability one. By contrast, in Rips's design, the critical transfer stimuli were equally similar to the nearest exemplars of the low- and high-variability categories, as illustrated in Figure 1.

2. In unidimensional identification, as the overall range of the stimuli is increased, the ability to discriminate between fixed pairs of stimuli tends to decrease. In addition, stimuli at the edges of the stimulus range are often discriminated with higher sensitivity than are stimuli located at the middle of the range. It is as if the perceptual space is "stretched" at its edges.

(Manuscript received May 3, 2001;
revision accepted for publication August 21, 2001.)