Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 5, 893–907

0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.5.893

The Effect of Category Variability in Perceptual Categorization

Neil Stewart and Nick Chater
University of Warwick

Exemplar and distributional accounts of categorization make differing predictions for the classification of a critical exemplar precisely halfway between the nearest exemplars of 2 categories differing in variability. Under standard conditions of sequential presentation, the critical exemplar was classified into the most similar, least variable category, consistent with an exemplar account. However, if the difference in variability is made more salient, then the same exemplar is classified into the more variable, most likely category, consistent with a distributional account. This suggests that participants may be strategic in their use of either strategy. However, when the relative variability of 2 categories was manipulated, participants showed changes in the classification of intermediate exemplars that neither approach could account for.

In this article, we consider the accounts of classification given by two successful models of categorization. Exemplar models (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) assume the categorization of a new exemplar is based on the similarity of the new exemplars to the representations of previously encountered exemplars stored in memory. An alternative is that probability distributions are used to represent categories and that these distributions are fitted by using the encountered exemplars. Classification of a new exemplar is based on the relative likelihood of belonging to each distribution. This alternative will be called the *distributional* approach (e.g., Ashby & Townsend, 1986).

The difference between these two accounts may be illustrated with a simple example in which the two accounts make qualitatively different predictions. Consider two categories (see Figure 1). The exemplars of one category may be more variable than the exemplars of the other category. If a *critical exemplar* exactly halfway between the nearest exemplars of the two categories is presented, it may be classified into either category. (The term *critical exemplar* is used to denote a novel test exemplar exactly halfway between the nearest neighbors of two categories.)

Exemplar models predict that the critical exemplar should be categorized as a member of the low-variability category more often than the high-variability category.[1] Intuitively, this is because the critical exemplar is, on average, nearer in perceptual space to the exemplars of the low-variability category and is therefore likely to be more similar to the exemplars of the low-variability category. Distributional models predict that the critical exemplar is more likely to be classified into the high-variability category. If the presumed distribution is Gaussian (see Figure 1), then the intermediate exemplar will typically, though not definitely,[2] be classified as a member of the high-variance category because the tight bunching of the low-variance exemplars means that the critical exemplar is more standard deviations from the mean of the low-variance category. (It is assumed here that the frequencies of each category are equal—in the experiments below, there is indeed no bias in favor of one category or the other.)

In summary, the exemplar and distributional models often make different predictions about the classification of a critical exemplar midway between the nearest exemplars from two categories differing in variability. We evaluate participants' performance on such a critical exemplar in Experiment 1. This idea is extended in Experiments 2 and 3, in which we investigate the effect of changing the relative variability of the two categories.

The effects of category variability on generalization have been addressed in two important studies: Rips (1989) and Fried and Holyoak (1984). Rips used a binary categorization with categories

Neil Stewart and Nick Chater, Department of Psychology, University of Warwick, Coventry, England.

This research was partly funded by a graduate assistantship awarded to Neil Stewart by the University of Warwick, partly by Economic and Social Research Council Grant R000239351 awarded to Gordon Brown, and partly by Biotechnology and Biological Research Council Grant 88/S09589 awarded to Evan Heit. Nick Chater was supported by European Commission Grant RTN-HPRN-CT-1999-00065. We thank Gordon Brown, Lewis Bott, Evan Heit, Koen Lamberts, Stian Reimers, and David Shanks for their helpful comments on this work.

Correspondence concerning this article should be addressed to Neil Stewart, Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: neil.stewart@warwick.ac.uk

[1] The exemplar model's exact predictions for the classification of the critical exemplar of course depends on the particular arrangement of exemplars. For example, if the high-variability exemplars just happen to be nearer to the critical exemplar, the opposite prediction would be made. However, if exemplars are randomly generated from normally distributed categories, this is unlikely to be the case.

[2] The reason it is not certain that the critical exemplar should be categorized as a member of the high-variability category more often than as a member of the low-variability category is because the critical exemplar is not equidistant between the means of the two categories (when this would always be the case). (It is worth pointing out here that if this were the case, then an exemplar model would be able to predict classification of the critical exemplar into the high-variability category as this category is most likely to have the nearest exemplar.) Rather, the critical exemplar is equidistant between the nearest neighbors of the two categories and is therefore nearer the mean of the lower variability category. Thus, the difference in variability between the two categories need be sufficiently large to counter the fact that the low-variability category has the nearer mean.

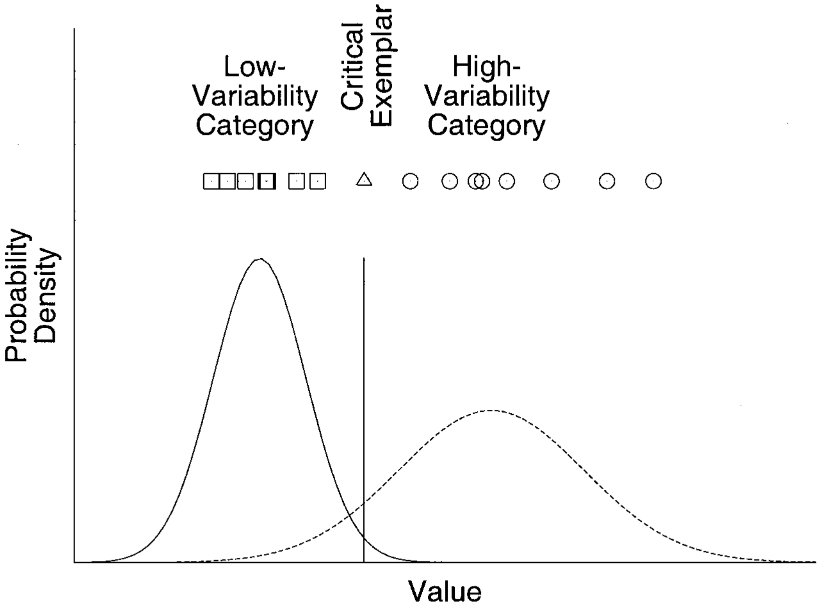

Figure 1. A one-dimensional example of two categories differing in variability. The exemplars of the low-variability category happen to take low values on the dimension (squares). The probability density function from which they were generated is represented by the solid line. The exemplars of the high-variability category take high values of the dimension (circles). The probability density function from which they were generated is represented by the dashed line. A critical example midway between the nearest examples of the two categories (triangle) is more likely to belong to the high-variability category but is more similar to examples of the low-variability category.

of differing variability to dissociate similarity and categorization judgments. Participants were presented with sentences giving information about an object's value on a single dimension. In one condition participants had to classify the object as a member of one of two available categories on the basis of this information alone. In another condition, participants were asked to choose the category to which the object was more similar.[3] The value of the object on the selected dimension was chosen to be halfway between the participant's estimates of the lowest value of the high-value category, and the highest value of the low-value category. Participants were told this is how the test value they were given was derived. Rips found that similarity decisions favored the low-variability category but that categorization decisions favored the high-variability category. Rips took the dissociation between similarity and categorization as evidence that categorization decisions were not based on similarity decisions. Empirical evidence from E. E. Smith and Sloman (1994) provided a pertinent boundary condition on this dissociation. They found that Rips's dissociation of categorization and similarity is only obtained under conditions that require verbal rationalization of the categorization decision.

Rips's (1989) study leaves open the question of the effect of category variability in perceptual categorization, the topic of the present article, for two reasons. First, Rips used familiar semantic categories to encourage participants to use prior knowledge from outside the experimental context. Such knowledge is not available for the kinds of abstract perceptual stimuli traditionally used in perceptual categorization experiments (although it may well be available for natural perceptual categories). Second, the effect that Rips described does not seem to be robust in conditions most analogous to those of a typical perceptual categorization task (where participants do not produce verbal protocols).

Fried and Holyoak (1984) have shown that participants are sensitive to the relative variability of perceptual categories. They found that participants classified some checkerboard patterns physically closer to the prototype (or mean) of a lower variability category as members of the high-variability category. Fried and Holyoak had predicted these findings with their category density model and interpret these findings as support for a distributional approach. However it is also consistent with exemplar-based categorization, as it is much more likely that there will be more exemplars from the high-variability category near the transfer checkerboard than exemplars from the low-variability category, simply because the checkerboards from the high-variability category are more scattered from their prototype. A second issue regarding Fried and Holyoak's interpretation is that their similarity estimate (i.e., the number of squares in common) may lead to incorrect assumptions about the representation of these checkerboard stimuli. To a first approximation it may be that the largest invariant chunk of a stimulus is learned as a feature (McLaren, 1997; Palmeri & Nosofsky, 2001; Stewart, 2001; Wills & McLaren, 1998). Because the low-variability category's exemplars vary less, this would lead to the creation of larger functional

[3] Note that participants were not asked for similarity ratings between two objects as is typical in predicting classification from similarity or identification (e.g., Nosofsky, 1986) but rather gave ratings of the similarity between an object and a category.

features for this category. If this were the case, then an exemplar equally distant between the two categories may indeed be more similar to the high-variability category simply because the probability of the presence of larger chunks used to represent the low-variability category is much lower than for the high-variability category.

What is needed is a category structure that allows the similarity and distributional models to be distinguished, even when memory for individual exemplars is allowed (as it is in the hugely successful exemplar models). Such a structure, illustrated in Figure 1, was offered above.

Modeling Sensitivity to Category Variability

To confirm the intuitive argument that exemplar and distributional models of categorization make opposite predictions, we examine two existing models of categorization in this section: the generalized context model (GCM; Nosofsky, 1986) and normal general recognition theory (Normal GRT; Ashby & Townsend, 1986).

First consider the predictions of the GCM. In the GCM, each encountered exemplar is represented as a point in a perceptual space. To classify a new exemplar, the similarity between the new exemplar and each stored exemplar is calculated. (Similarity is a decreasing function of the distance between exemplars in perceptual space.) Similarities are then summed for each category. Luce's (1959) choice rule is used on the summed similarities to calculate the probability that the exemplar is classified into a given category. Figure 2A plots the probability that the exemplar is classified into the high-variability category as a function of the exemplar's location. This function is referred to as the *generalization gradient*. The different gradients correspond to different values of the generalization parameter, *c*. For broad generalization (i.e., small *c*) the similarity of a given exemplar to more distant exemplars will be larger than for narrow generalization (i.e., large *c*). Thus when generalization is narrow, the generalization gradient is steeper. Provided the exemplars are appropriately arranged, the model predicts that the critical exemplar is most likely to be classified into the low-variability category for any value of the generalization parameter. The predictions here are for the GCM with a Gaussian function ($q = 2$) relating similarity to distance. The predictions of the GCM with an exponential similarity function ($q = 1$) do not differ qualitatively.

We illustrate the distributional approach by using Normal GRT. Normal GRT is an extension of standard GRT. In standard GRT each exemplar is represented by a normal distribution in perceptual space. Thus standard GRT would make similar predictions to the GCM, as each model assumes (some) memory for each exemplar. In contrast, in Normal GRT each category, rather than each exemplar, is represented by a single normal distribution. Ashby (1992) made the strong assumption that many natural categories can be represented by a normal distribution even when the true distribution is not normal. In Normal GRT, the category exemplars are used to infer a population mean and variance for the normal representation for each category. An optimal decision bound is then calculated that divides the perceptual space into regions for each category, so that all the exemplars represented by points in the same region are most likely to belong to a common category. In the one-dimensional case for two categories of unequal variance, the optimal decision bound will be a pair of points, with the lower variability category in between the two points and the higher variability category outside the pair. Perception is assumed to be noisy in GRT. Thus an exemplar near the decision bound may sometimes be perceived to fall on one side of the bound and sometimes on the other. To apply Normal GRT to the category

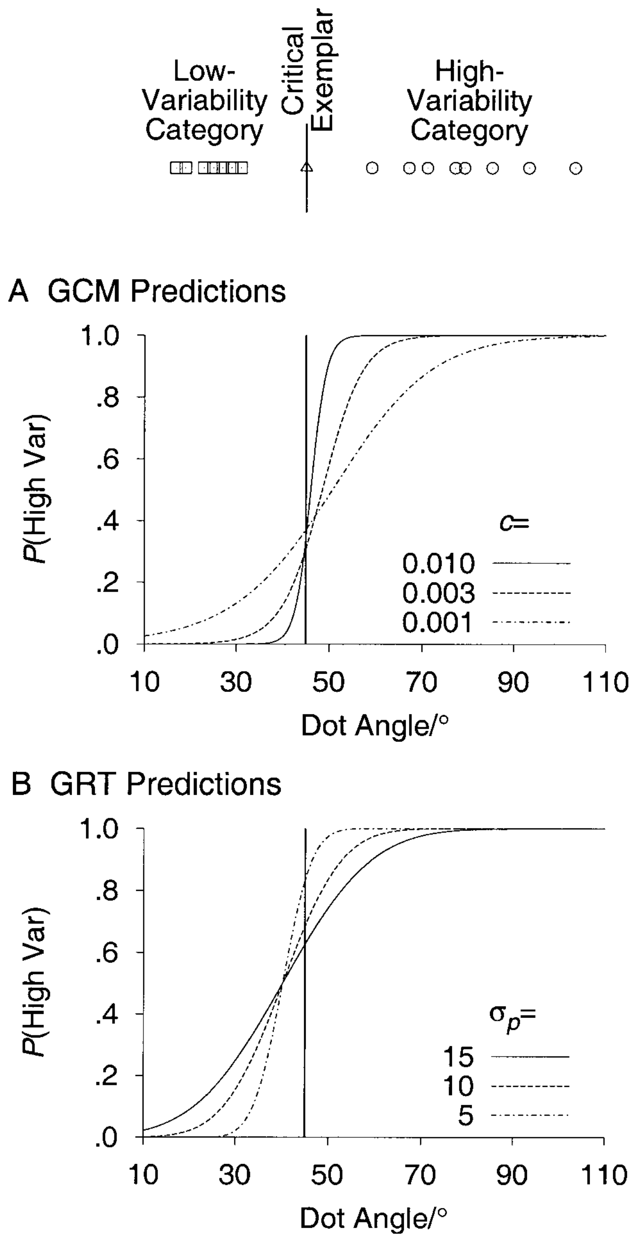

Figure 2. Predictions for the probability of a high-variability-category response plotted as a function of the stimulus value for the stimuli used in Experiment 1. The category structure is illustrated along the top of the figure, with one category more variable than the other. A: Predictions for the generalized context model (GCM). The three lines correspond to different values of the generalization parameter, *c*. B: Predictions for normal general recognition theory (GRT). The three lines correspond to different levels of perceptual noise, which is assumed to be normally distributed with standard deviation σ_p.

structure for Experiment 1, we used the eight exemplars for each category to generate an estimate of the population mean and variance of the normal distribution from which the exemplars were generated.[4] The optimal decision bound was then calculated. The exact predictions for classification of exemplars near the decision bound depend on the level of perceptual noise (σ_p). Following Ashby and Townsend (1986), we assumed the perceptual noise to be Gaussian. Figure 2B illustrates three generalization gradients. The less noise, the steeper the generalization gradient. Crucially though, the level of noise changes the slope of the generalization gradient but does not alter the location of the optimal decision bound.

In summary, for a critical exemplar that lies exactly between the nearest neighbors of two categories that differ in variability, the GCM often predicts this critical exemplar is more likely to be classified into the low-variability category (independent of the amount of generalization), and Normal GRT predicts that the critical exemplar is more likely to be classified into the high-variability category (independent of the amount of perceptual noise).

Experiment 1

Experiment 1 was designed to discriminate between exemplar-based classification and distribution-based classification by using a category structure as described above. In one condition participants were given a hint telling them that the two categories differed in variability. E. E. Smith and Sloman's (1994) replications of Rips's (1989) study suggest that participants categorize stimuli into the high-variability category only when their verbal protocols show awareness of a difference in variability between the two categories. The hint here was included to see what effect knowledge of the variability difference might have on participants' classification. The method of presentation of the exemplars was manipulated as an additional between-participants factor. During the learning phase, exemplars were either presented sequentially or simultaneously. We hypothesized that simultaneous presentation should make the difference in the variability of the categories more salient.

Method

Participants. Sixty-four undergraduate students from the University of Warwick participated for course credit.

Design. Participants performed three binary categorization tasks. There was a separate stimulus set for each of the three tasks. After learning 16 training exemplars, participants classified a critical exemplar that fell halfway between the nearest exemplar of the low-variability category and the nearest exemplar of the high-variability category. They then classified two further verification exemplars, one from each category, before moving on to the next classification. There were two between-participants factors: (a) simultaneous or sequential presentation of training exemplars and (b) whether participants were given a hint that one category was more variable than the other.

Stimuli. An example stimulus set is shown in Figure 3. The stimuli used in this experiment were outline circles each with a single solid dot somewhere on their circumference. The diameter of the circle subtended approximately 2° of visual angle. The stimuli varied only in the position of the dot around the circumference; this position was diagnostic of category membership. Pilot studies used the position of the dot on a straight line, but the performance of many participants was consistent with their reports of using a rule, such as whether the dot was more or less than halfway along the line, to make their decision. The stimuli here were chosen so that use of rules like this (e.g., using horizontal, vertical, or diagonal diameters as decision bounds) should not be possible.

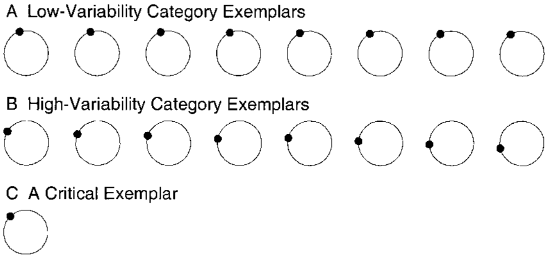
A Low-Variability Category Exemplars
B High-Variability Category Exemplars
C A Critical Exemplar

Figure 3. An example of a stimulus set from Experiment 1.

For each participant, for each category, eight exemplars were generated from a normal distribution. The low-variability category distribution had a standard deviation of 11°, and the high-variability category had a standard deviation of 28°. There was a gap of 56° between the nearest exemplars of each category, with the critical exemplar lying exactly in the center of this gap. To ensure the gap between the nearest neighbors of each category was constant for all participants, the means of the categories needed to be adjusted slightly for each participant. The critical exemplar was in the 45° position for the first task, the 135° position for the second task, and the 225° position for the third task (with 0° being at the 12 o'clock position and angle increasing counterclockwise). The relative position of the low- and high-variability categories was counterbalanced across participants.

Because the exact predictions of the GCM and normal GRT depend on the particular distribution of exemplars, all of the stimulus sets were modeled to check that the critical exemplar was indeed more similar to the low-variability category but more likely to belong to the high-variability category. This was always the case.

Apparatus. For the sequential presentation condition, stimuli were displayed on a 14-in (36-cm) Apple Macintosh Color Display and responses were collected by using labeled keys on a standard qwerty keyboard (the keys *A* to *J*, inclusive, were labeled *A*, *B*, *C*, *yes*, *D*, *E*, and *F*, respectively). For the simultaneous presentation condition, stimuli were presented in a 210 × 297 cm booklet and responses written into the booklet.

Procedure. The experiment began with instructions telling participants they would do three categorization tasks, one after the other. Participants in the hint condition received further instructions telling them that one (but not which) category was allowed a greater spread of dots than the other. They were instructed to try to identify the category that had the greater spread of dots during the experiment.

In the sequential presentation condition, each trial began with a ready prompt. When a participant pressed *yes*, there was a 1.5-s blank screen before a circle with a dot appeared on the screen for 1 s. Participants responded as quickly and accurately as they could from stimulus onset. The assignment of category labels to the high- and low-variability categories was counterbalanced across participants. After 1 s, the screen was cleared, whether the participant had responded or not. After the participant responded, the correct answer was displayed on the screen for 1.5 s, followed by a 1.5-s blank screen before the next trial began. The feedback for the critical exemplar was random, so participants' attention was not drawn to the special status of the critical exemplar (which might have affected performance on later stimulus sets).

[4] In fact, because perception is assumed to be noisy, this method only provides the best estimate of a participant's hypothesized mean and variance.

The same stimuli were used for the simultaneous presentation condition, which began with presentation of the first stimulus set. Each set of eight exemplars belonging to the same category was arranged in a row, inside a rectangle, together with the category label. The two sets were placed one above the other. The placement of the low- and high-variability categories at the top and bottom of the page was counterbalanced across participants, as was the assignment of labels to categories. Within a set, the exemplars were arranged in the same (random) rank order for all participants to ensure that if the order of the exemplars on the page affected the salience of the variability, then it would be held constant across conditions. Participants studied the sheet of exemplars for 1 min, and then it was removed from sight. The critical exemplar was then presented in the center of a new piece of paper. Participants circled the category label to which they thought the exemplar belonged. This was repeated with the verification exemplars.

Results

Data were collapsed across all three stimulus sets. For the sequential condition, the mean training proportion correct was high (no hint: mean proportion correct = .81, *SE* = .02; hint: mean proportion correct = .79, *SE* = .02) and did not differ between the hint and no-hint conditions, $t(31) = 0.85$, $p > .05$. No training data were collected in the simultaneous presentation condition. However, performance can be compared across the simultaneous and sequential conditions by using the verification trials. Verification performance averaged across all conditions was high (mean proportion correct = .93, *SE* = .02). A two-way analysis of variance (ANOVA) (Hint × Presentation) revealed no effect of hint, $F(1, 60) = 1.56$, $p > .05$, no effect of presentation, $F(1, 60) = 0.39$, $p > .05$, and no significant interaction, $F(1, 60) = 0.00$, $p > .05$. In summary, knowledge that the two categories differed in variability did not facilitate category learning and neither did presentation method.

Of most interest is performance on the critical exemplar. Table 1 shows the proportion of high-variability responses averaged across all three critical exemplars. A two-way ANOVA (Hint × Presentation) was run. Simultaneous presentation increased the proportion of high-variability responses, $F(1, 60) = 18.56$, $p < .05$, as did giving a hint that the two categories differed in variability, $F(1, 60) = 5.96$, $p < .05$. There was no significant interaction, $F(1, 60) = 0.52$, $p > .05$. Planned *t* tests were run to see which means differed significantly from chance performance of .5. For the sequential presentation conditions, the proportion of high-variability responses was significantly below chance for both the hint condition, $t(15) = 7.31$, $p < .05$ and the no-hint condition, $t(15) = 13.17$, $p < .05$. For the simultaneous presentation condition, the proportion of high-variability responses was not significantly different from chance for the no-hint condition, $t(15) = 0.13$, $p > .05$, but was significantly above chance for the hint condition, $t(15) = 3.61$, $p > .05$.

Table 1
Mean Proportion of High-Variability Responses in Experiment 1, Split by Hint and Presentation Method

	Presentation	
Condition	Sequential	Simultaneous
Hint	.37 (.09)	.74 (.07)
No hint	.25 (.06)	.51 (.08)

Note. Numbers in parentheses are standard errors of the means.

Discussion

In this experiment a critical exemplar lying midway between the nearest exemplars of two categories differing in their variability was significantly more likely than chance to be classified as belonging to the lower variability category when training exemplars were presented sequentially. This pattern of classification is consistent with the prediction of exemplar models—that is, that the critical exemplar should be classified into the more similar category. When training exemplars were presented simultaneously, participants were significantly more likely to classify exemplars into the high-variability category than when they were presented sequentially. When participants were given a hint that the two categories differed in variability, they were significantly more likely to classify the critical exemplar into the high-variability category. In combination, simultaneous presentation and hint caused participants to classify the critical exemplar into the high-variability category more often than chance, consistent with the predictions of distributional models—that is, that the critical exemplar should be classified into the category most likely to have generated it. However, both models were originally designed to explain sequential categorization performance, and the data collected under sequential presentation conditions here support an exemplar account rather than a distributional account.

Note that this experiment provides no evidence that the critical exemplar was midway between the nearest exemplars of the two categories in participants' psychological space. However, it is at least reasonable to assume that the psychological-space critical exemplar must be in the region of the test critical exemplar that was actually presented. Therefore given the large sizes of the effects of presentation and hint, even if the psychological-space critical exemplar does not coincide precisely with the physical-space critical exemplar, its classification would also be strongly influenced by these factors.

There are two possible alternative accounts of these findings. The first is that changing the method of presentation and providing a variability hint alters the representation of the categories that participants form, rather than altering the classification strategy they use. Consider how this account would work if participants were using an exemplar strategy in all conditions of this experiment. The shift to classification of the critical exemplar into the high-variability category with simultaneous presentation and hint would have to be explained as exemplars of the high-variability category being closer in perceptual space to the critical exemplar under these conditions compared with the sequential presentation and no-hint conditions. However, the switch from sequential presentation and no hint to simultaneous presentation and hint was intended to have exactly the opposite effect (i.e., to draw attention to the variability difference). Thus although this alternative account remains a possibility, it does not seem plausible. However, consider how the changing representation account would explain these data if participants were using a distributional strategy throughout the experiment. In this case, switching from sequential to simultaneous presentation and providing the variability hint should allow participants to assign a larger variability distribution to the more variable category in the simultaneous hint condition rather than the sequential no-hint condition. This leads to the prediction that the critical exemplar will be classified into the high-variability category most often in the simultaneous hint condition. In the sequential condition, when the difference in variability is not salient, participants might assume that the two categories had equal variance. Thus as the critical exemplar is nearer to the mean of the low-variability category, the distributional account predicts that it should be classified into this category most often. Both of these predictions are consistent with these data.

The second alternative account of these data is that the response bias changes systematically between these conditions. To account for these data, the bias for the high-variability category would have to have increased when presentation was switched from sequential to simultaneous presentation and a hint was provided. We return to this possible account below.

Sensitivity of Exemplar and Distributional Models to Changes in the Relative Variability of Categories

In Experiment 2, we investigate how changing the relative variability of two categories should affect the classification of intermediate exemplars. (The term *intermediate exemplars* denotes any exemplars between the two categories, in contrast to the use of the term *critical exemplar*.) The category structures used are illustrated in the top panel of Figure 4 and are described in detail in the *Design and stimuli* section of Experiment 2. The stimuli were rectangles or ellipses, defined by their height and width. One pair of categories had standard deviations in the ratio of 1:2; the other pair had standard deviations in the ratio of 1:4. Across conditions, the low-variability categories had equal means. The high-variability categories also had equal means. Finally, the distance between the nearest neighbors of each category was constant across the 1:2 and 1:4 conditions.

Given the category representation of the Normal GRT, it seems likely that this model would be sensitive to differences in the relative variability of two categories. This is indeed the case. All the categories are represented using simple covariance matrices ($\Sigma = \sigma^2 I$) because of the symmetrical nature of the categories. In general, with two bivariate normal categories differing in covariance matrix the decision bound is quadratic (Ashby, 1992, p. 460). Here we modeled performance for stimuli lying on the line between the two category means (i.e., *height* = *width*). As in modeling for the category structure used in Experiment 1, the perceptual noise changes the shape of the generalization gradient but does not bias the decision bound (i.e., the point at which a stimulus is equally likely to be classified into either category) one way or the other. Of interest here is the comparison of gradients for the 1:2 and 1:4 conditions. One generalization gradient for each condition is shown in Figure 5A. (The level of perceptual noise is assumed constant across both structures, $\sigma_p = 10$.) As the difference in variability between the two categories is increased the decision bound moves nearer to the low-variability category.

The variances of each category were chosen to keep the distance between the nearest exemplars of each category constant across the 1:2 and 1:4 conditions. This allows an alternative comparison in which the classification of intermediate exemplars that are the same distance from the nearest neighbor of the low-variability category is contrasted (i.e., with the same coordinates, relative to the nearest neighbors). Because the distance between the nearest neighbors of each category is held constant across the 1:2 and 1:4 conditions, intermediate exemplars that are equally distant from the nearest neighbor of the low-variability category across conditions must also be equally distant from the nearest neighbor of the high-variability category across conditions. For comparison of exemplars with either the same absolute coordinates (see Figure 5A), or the same coordinates relative to the nearest neighbors (see Figure 5C), each exemplar is always predicted to be more likely to be classified into the high-variability category in the 1:4 condition compared with the 1:2 condition. This is always true for any level

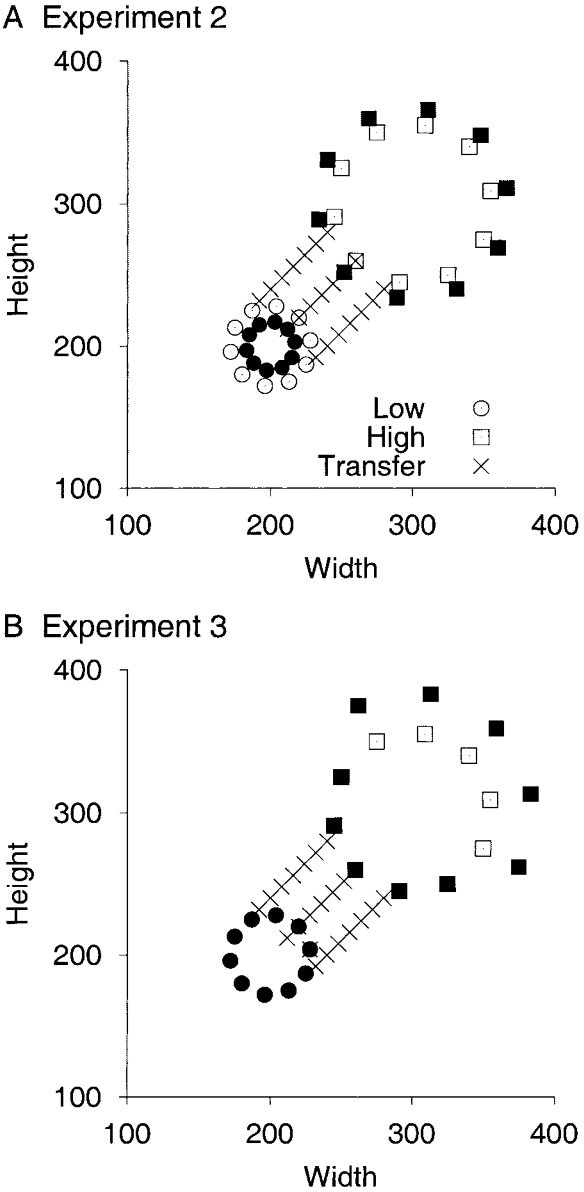

Figure 4. The arrangement of exemplars in Experiments 2 and 3. The open shapes represent the 1:2 condition that is used in Experiments 2 and 3. A: For Experiment 2, the solid shapes represent the 1:4 condition. B: For Experiment 3, the solid shapes represent the 1:2 Expanded condition (and cover all of the low-variability category exemplars and half of the high-variability category exemplars from the 1:2 condition.)

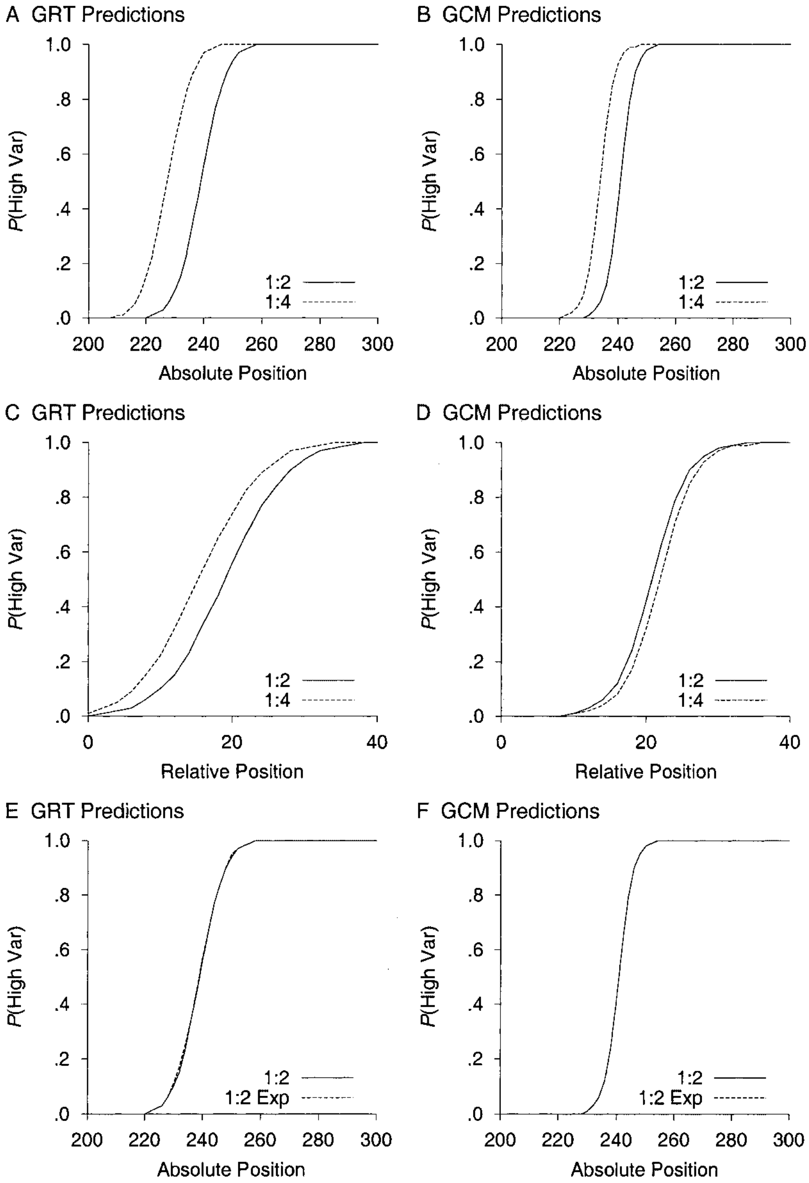

Figure 5. Predictions for the probability of a high-variability-category (High Var) response plotted as a function of the stimulus width (or height) for Experiments 2 and 3. The label "absolute position" refers to the actual size of exemplars. The label "relative position" refers to the size of the exemplar compared to the nearest exemplar of the low- (or high-) variability category. A: Predictions of normal general recognition theory (GRT) for Experiment 2 ($\sigma_p = 10$). B: Predictions of the generalized context model (GCM) for Experiment 2 ($q = 2$, $r = 2$, $c = 0.05$). C: Normal GRT predictions for Experiment 2 shown in Panel A plotted as a function of relative position, rather than absolute position. D: The GCM predictions for Experiment 2 shown in Panel B plotted as a function of relative position, rather than absolute position. E: normal GRT predictions for Experiment 3 ($\sigma_p = 10$). F: The GCM predictions for Experiment 3 ($q = 2$, $r = 2$, $c = 0.05$). In Panels E and F the gradients for the two conditions are almost exactly coincident. Exp = expanded.

of perceptual noise because perceptual noise alters only the slope of the generalization gradient and not the location of the decision bound.

The generalization gradients predicted by the GCM for the two category structures are also shown in Figure 5B, with the generalization parameter held constant ($c = 0.05$) across the two structures. The predictions here are for the GCM with a Euclidean distance metric ($r = 2$) and a Gaussian similarity function ($q = 2$); however, the pattern of the predictions is the same for a city block distance metric ($r = 1$) and exponential similarity function ($q = 1$). The predictions of the GCM are similar to those of Normal GRT. In the 1:4 condition, the high-variability category's exemplars are nearer, and the low-variability category's exemplars are further away, from a given intermediate exemplar, compared with the 1:2

condition. Therefore exemplars intermediate between the two categories are more likely to be classified as members of the high-variability category in the 1:4 condition than in the 1:2 condition. However, when the generalization gradients are measured relative to the two nearest neighbors, this is no longer true (see Figure 5D). When exemplars an equal distance from the nearest neighbor of the low-variability category in each condition are compared, classification into the high-variability category is more likely in the 1:2 condition because the second nearest neighbors of the high-variability category are nearer in the 1:2 condition than in the 1:4 condition and the second nearest neighbors of the low-variability category are further away in the 1:2 condition than in the 1:4 condition. (Note that this follows because (a) the exemplars of the high-variability category are more spread out in the 1:4 condition than in the 1:2 condition and (b) the low-variability category exemplars are less spread out in the 1:4 condition than in the 1:2 condition.) This prediction is the opposite prediction to Normal GRT. For these category structures it is trivial to prove that this prediction is true for all amounts of generalization.[5]

Experiment 2

In Experiment 2 generalization gradients were obtained for participants after training on both the 1:2 and 1:4 conditions. Experiment 2 sets out to find which model describes the behavior of participants, both at the level of across-participant averages and also at the level of individual participants. It is important to consider performance at the level of individual participants, particularly in view of the demonstration by Maddox (1999; see also Ashby, Maddox, & Lee, 1994) that data averaged across participants might not reflect individual participant data, especially when large individual differences exist. Using Monte Carlo simulation, Maddox generated data sets from either GRT or from the GCM. When the GCM was the correct model, averaging had little effect. However, when GRT was the correct model and therefore perfectly described the generated data, averaging led to a better fit for the GCM. This implies that averaging the data alters the qualitative structure of the data. Thus, averaged data should not be used to compare the two models, as averaging the data biases the result in favor of the GCM.

Method

Participants. Thirty-two undergraduates from the University of Warwick participated for course credit or payment of £5 (U.S. $7.39).

Design and stimuli. Each participant completed two categorization training and transfer tasks. In the training stage, participants learned to categorize stimuli that varied in height and width into one of two categories, with trial-by-trial feedback. In the transfer stage, participants classified old training exemplars and new transfer exemplars without feedback.

The tasks differed in the category structure used (see Figure 4A). Both category structures had two categories, one with a mean of (200,200) and the other with a mean of (300,300) in units of pixels. The 10 exemplars of each category were arranged in a circle around each mean. In the 1:2 condition the low-variability category was half as variable as the high-variability category (standard deviation of 20.0 vs. 40.0 on each dimension), and in the 1:4 condition the low-variability category was about four times less variable than the high-variability category (standard deviation of 12.7 vs. 50.2 on each dimension). In the transfer stage, additional exemplars intermediate in height and width between the two categories were included to measure the generalization gradient.

The order of learning the 1:2 and 1:4 tasks was counterbalanced across participants. To minimize carry-over effects, in one condition stimuli were rectangles of varying height and width and in the other condition stimuli were ellipses of varying height and width. The assignment of shape to condition was counterbalanced across participants. The assignment of labels to categories was also counterbalanced. Finally, the assignment of variability to the category of either small or large stimuli was also counterbalanced. That is, for half the participants, the category with the smaller stimuli was the less variable category (as in Figure 4A), and for the other half, the category with the larger stimuli was the more variable category (the mirror image of Figure 4A, about the line $height + width = 500$).

It is not always the case that a category structure in psychological space reflects the structure of the category in the experimenter's choice of physical space (e.g., Palmeri & Nosofsky, 2001). A separate experiment, not reported here, was run in which pairwise similarity judgements were obtained for the stimuli used. The individual differences multidimensional scaling model (Carroll & Wish, 1974; Shepard, 1980) was used to derive solutions for the 1:2 and 1:4 conditions. Examination of the solutions confirmed that the ratio of the mean interexemplar distance within each category was greater for the 1:4 condition than for the 1:2 condition. This supports the key assumption in this experiment—that the representation of one category was indeed more variable than the other, and further, that the difference in variability was greater in the 1:4 condition than in the 1:2 condition.

There is some debate on the nature of the psychological representation of rectangles (e.g., Feldman & Richards, 1998; Krantz & Tversky, 1975; Macmillan & Ornstein, 1998; Monahan & Lockhead, 1977). Krantz and Tversky (1975) suggested that dimensions of area ($a = h \cdot w$) and shape ($s = h/w$) may be more appropriate than height (h) and width (w). Further, the space may also be subject to Weberian compression for larger heights and widths. However, under transformation to a-s space, $\log(h)$-$\log(w)$ space and $\log(a)$-$\log(s)$ space, the qualitative properties outlined in the previous paragraph remain unaltered.[6]

Apparatus. Stimuli were displayed on a 14-in (36-cm) Apple Macintosh Color Display. Responses were collected using labeled keys on a standard qwerty keyboard. The keys *Z* and *X* were labeled *A* and *B* respectively.

Procedure. Each trial started with the presentation of a stimulus until the participant responded. Feedback was given on the screen for 1,500 ms. The feedback was the correct category label, presented as a letter (*A* or *B*) 50 pixels high below the stimulus. The stimulus remained on the screen until the end of the feedback. The screen was then blank for 500 ms before the next trial began automatically. The sequence of 100 trials comprised five repetitions of the 20 training exemplars. In each repetition, the trials were in a random order. The 328 transfer trials comprised eight repetitions of 41 exemplars. Of the 41 exemplars, 20 were the old training exemplars; the remaining 21 transfer exemplars were novel exemplars located in between the two categories in height–width space. Within each repetition, the 41 exemplars were displayed in a random order. The structure of a trial was the same as in training, except the feedback was omitted. After a participant had responded, the screen was cleared, and the next trial began after a 500-ms pause. When participants had completed the first categorization task, they moved on to a second task, which was the same as the first

[5] Proof follows by writing out, for each category structure, the expression for the probability that a given intermediate exemplar will be classified into the high-variability category according to the GCM and then showing that this value is greater for the 1:4 condition than for the 1:2 condition for all values of c, when exemplars equally distant from the nearest neighbors of either category are compared.

[6] We thank Thomas S. Wallsten for drawing these alternative potential representations to our attention.

except that the category structure was swapped, as was the type of shape. No instruction that the categories differed in variability was given.

Results

Average results. Participants were very accurate in their training classifications. On average, the mean proportion of correct responses in training was .91. A six-way ANOVA (Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse × Condition × Category) was run to check that none of the counterbalanced factors or the category structure affected training performance. There was a significant effect of category mean and variance assignment, corresponding to a slight improvement in accuracy when the category with the low mean had the lower variance (.94 vs. .91), $F(1, 16) = 7.03$, $p < .05$. This effect was not found in transfer. There were no other significant main effects, largest $F(1, 16) = 1.42$, $p = .25$.

Performance on old training exemplars was also excellent during transfer. The proportion of high-variability-category responses to old training exemplars is shown in Table 2. A six-way ANOVA (Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse × Condition × Category) revealed a main effect of category, $F(1, 16) = 6.54$, $p < .05$. Although performance was high on training exemplars in test, exemplars of the low-variability category were classified slightly less accurately than exemplars of the high-variability category (mean proportion correct = .89 versus .96). There were no other significant main effects, largest $F(1, 16) = 2.32$, $p > .05$. This indicates that no counterbalanced factor had a significant effect on old training exemplar classification in transfer.

It is the performance on the new transfer exemplars that is of interest. The responses given to each of the 21 new transfer exemplars are collapsed into seven sets, so that responses to stimuli whose projections onto the line *height* = *width* coincide were in the same set. Figure 6A shows a plot of the proportion of high-variability responses given to stimuli in each of the seven sets as a function of their size. Figure 6A can therefore be thought of as showing a generalization gradient. A six-way ANOVA (Condition × Stimulus Set × Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse) was run. In both the 1:2 and 1:4 conditions, the proportion of high-variability responses to test exemplars increased as the location of the test exemplar moved toward the high-variability category, $F(6, 96) = 185.77$, $p < .05$ (Huynh–Feldt $\varepsilon = .82$). In the 1:4 condition the proportion of high-variability responses was higher than for the 1:2 condition for every set of test stimuli, $F(1, 16) = 10.52$, $p < .01$. There was no significant interaction between stimulus and condition, $F(6, 96) = 1.67$, $p > .05$ (Huynh–Feldt $\varepsilon = 1.00$). There were no other significant main effects, largest $F(1, 16) = 1.06$, $p > .05$, showing that none of the counterbalanced factors affected responding significantly.

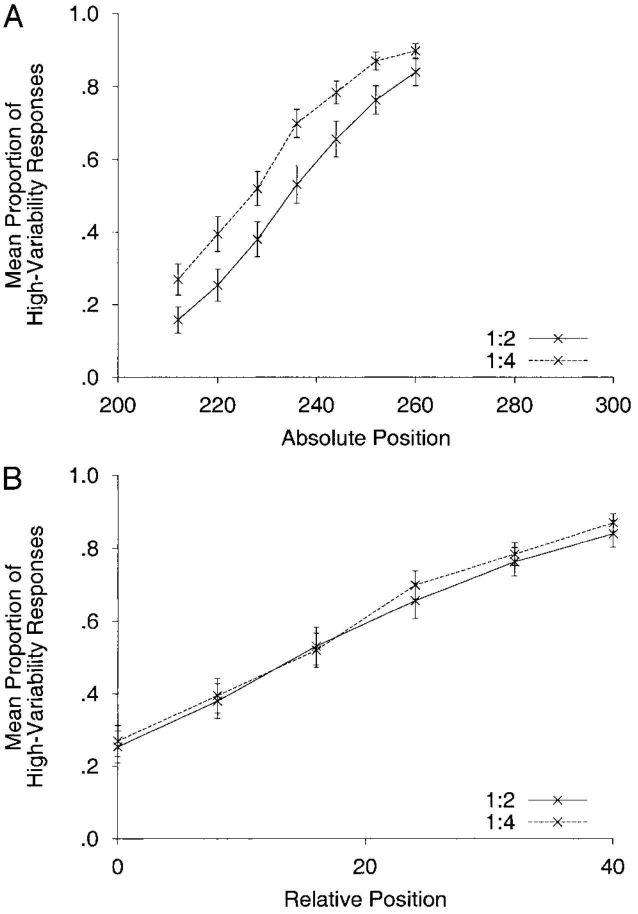

Figure 6. The results of the transfer stage of Experiment 2. In Panel A, the results are plotted as a function of absolute position; in Panel B, the same results are plotted as a function of relative position.

Table 2
Mean Proportion of High-Variability Responses to Old Training Exemplars in Test for Experiment 2

	Condition	
Category	1:2	1:4
Low variability	.11 (.03)	.12 (.03)
High variability	.95 (.01)	.96 (.01)

Note. Numbers in parentheses are standard errors of the means.

By analyzing the results as above, we compared classification of exemplars that are equally distant from the mean of the low-variability category (or the mean of the high-variability category—the two comparisons are equivalent given the category structures used here) across the 1:2 and 1:4 conditions. However, an exemplar that is equally distant from the low-variability category mean in the 1:2 and 1:4 conditions is not equally distant from the nearest exemplar of the low-variability category in both conditions. The following analysis compares exemplars that are equally distant from the nearest exemplar of the low-variability category across the two conditions. (As the distance between the nearest neighbors of each category was the same for both conditions, it does not matter whether distance is measured relative to the position of the low-variability category's nearest exemplar or to the high-variability category's nearest exemplar.) Such a comparison is shown in Figure 6B. (If one shifts the 1:2 data in Figure 6A one step to the left, one obtains Figure 6B.) Another six-way ANOVA (Condition × Stimulus Set × Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse) was run. Unsurprisingly, as before, as the location of the test exemplar got nearer the exemplars of the high-variability category, the proportion of high-variability responses increased, $F(5, 80) = 170.01$, $p < .05$ (Huynh–Feldt $\varepsilon = .87$). However, now that position is measured relative to the nearest neighbors of the two categories, there is no difference between the generalization gradients for the two conditions, $F(1, 16) = 0.23$, $p > .05$. There was no stimulus by condition interaction, $F(5, 80) = 0.41$, $p > .05$ (Huynh–Feldt $\varepsilon = 1.00$). There were no other significant main effects, largest $F(1, 16) = 1.19$, $p > .05$, showing that none of the factors counterbalanced across participants affected responding significantly.

Individual participant results. When generalization gradients were calculated for individual participants, many participants showed very different gradients for the two conditions. The results averaged across participants did not represent individual performance well. Even when the effect of nearest neighbors was controlled, many participants showed a difference in gradients. Further, for many of these participants, the change was larger than would be expected by chance. A chi-square analysis was performed for each participant, with the trial as the unit of analysis. A 2 (variability condition) × 2 (response) contingency table was constructed for each participant containing the frequencies of low- and high-variability responses in each condition summed across transfer exemplars that were equally distant from the nearest neighbors of each category. A chi-square statistic was calculated on the basis of the hypothesis that there should be no difference in the proportion of high-variability responses between the two conditions. Yates's continuity correction was not used, as there is no reason to expect constant marginal totals, and the expected frequencies were large (Howell, 1997, p. 146). As the assumption that the response on each trial is independent of the response on any other trial is unlikely to be true, the statistic was deflated to account for trials being nonindependent (Altham, 1979; see also Tavaré & Altham, 1983). Thirteen of the 32 participants showed a significant difference between their responding in the two conditions, 7 increasing and 6 decreasing their proportion of high-variability responses as the difference in variability between the two conditions increased. The probability of obtaining 13 or more significant differences (i.e., $p < .05$) by chance is 1.72×10^{-9}, assuming that the number of significant results is binomially distributed ($n = 32$, $p = .05$).

Discussion

Averaged across participants, when the difference in variability between two categories was increased, the proportion of high-variability responses to intermediate exemplars increased. This result is consistent with the predictions of the GCM and of Normal GRT. Of interest here is the result when the presence of nearest neighbors was taken into account. This was done by comparing exemplars that were equally distant from the nearest neighbor of the low-variability category across the two conditions. Averaged across participants, the generalization gradients for the two conditions were virtually identical. This is inconsistent with the predictions of Normal GRT but is consistent with those of GCM (when the amount of generalization is small). However, the individual participant data were not well described by the average results.

A significant minority of participants showed a significant difference in their relative position generalization gradients between the two conditions. For about half of this minority, the relative position generalization gradient was shifted toward the low-variability category in the 1:2 condition compared with the 1:4 condition, consistent with the predictions of the GCM. For the other half, the shift was in the opposite direction, consistent with GRT. The majority of participants showed no significant change in relative position generalization gradient. Thus at the level of individual participants, some participants were behaving as if they were using an exemplar strategy and not a distributional strategy, and some participants were behaving as if they were using a distributional strategy and not an exemplar strategy. These data then do not provide support for one model over the other, and instead, at least for a significant minority of participants, challenge both models.

There is an alternative explanation: either the perceptual spaces formed, or the response biases used, in each condition fluctuated randomly for each participant.[7] Thus, participants may all be using the same categorization strategy, and the differences in the change in generalization gradient between participants may instead be due to random fluctuations. This is consistent with the observation that for those participants who showed a significant difference in relative position generalization gradient, half showed a shift in one direction and half showed a shift in the other direction. We address the possibility of such random fluctuations in Experiment 3.

Experiment 3

In Experiment 3, we used the 1:2 condition described above and a new condition. This new condition, 1:2 Expanded, differs only slightly from the 1:2 condition—in the 1:2 Expanded condition the five exemplars of the high-variability category that are furthest from the low-variability category are moved to even more extreme points (see Figure 4B). These two conditions are designed to allow the exemplar and distributional models to be further tested. Figure 5F shows the generalization gradients predicted by the GCM (Gaussian similarity function, Euclidean distance metric, $c = 0.05$) for the two conditions. The gradients almost exactly coincide. This is true for the range of c parameters that produces acceptable accuracy for the training exemplars (i.e., greater than 80% accuracy—participants in fact performed at about 90% accuracy). This can be explained intuitively as follows. When classifying exemplars from one category, the amount of generalization must be small enough to prevent generalization to exemplars in the other category. When the generalization is this small, the distant exemplars of the high-variability category in both category structures have only an infinitesimal level of similarity to the intermediate exemplars and thus have a negligible role in the classification of the intermediate exemplars. Therefore, moving these distant exemplars to even more distant locations in perceptual space should have no effect. In summary, if the GCM is to predict realistic accuracy for classification of old training exemplars, it is con-

[7] We thank Robert M. Nosofsky for suggesting this hypothesis as an alternative explanation.

strained to predict no difference between classification of intermediate exemplars between the 1:2 and 1:2 Expanded conditions.

As described above, the distant exemplars of the high-variability category in the 1:2 Expanded structure were moved to a distant location. This movement causes the high-variability category mean to move to a slightly more distant location in space. Modeling with Normal GRT for the 1:2 and 1:2 Expanded conditions shows that the effect of the increase in variability is almost exactly canceled out by this movement of the mean (see Figure 5E). The two generalization gradients are almost identical and are certainly empirically indistinguishable. Normal GRT then makes the same prediction as the GCM—that is, that there should be no difference in the generalization gradients for the two conditions.

Both the exemplar and distributional approaches were unable to predict the large variation between individuals demonstrated in Experiment 2. However, if some participants are assumed to apply an exemplar approach and some, a distributional approach, this variation might be explained. Our aim for Experiment 3 was to discriminate between these two possibilities. As demonstrated above, the GCM and Normal GRT predict no difference between the generalization gradients for the 1:2 and 1:2 Expanded conditions. However, the category structures used here are very similar to those used in Experiment 2, so there is good reason to expect replication of the large individual differences.

Method

This experiment differs from Experiment 2 only in the category structures used.

Participants. Thirty-two undergraduates from the University of Warwick participated for course credit or payment of £5 (U.S. $7.39). No participant had taken part in any other experiment in this study.

Stimuli. The stimuli in the 1:2 condition were the same as in Experiment 2. A new category structure, 1:2 Expanded (see Figure 4B), replaced the 1:4 structure.

As in Experiment 2, a separate multidimensional scaling experiment (not presented here) was run. Using the same method as described in Experiment 2, the ratio of the recovered mean within-category interexemplar distances was greater in the 1:2 Expanded condition than in the 1:2 condition. The similarity between the intermediate exemplars and the far exemplars of the high-variability category in both the 1:2 and 1:2 Expanded conditions (when calculated as in the GCM) was negligible compared with the similarity to other training exemplars, for *c* parameters large enough to produce acceptable accuracy on the old training exemplars in test. This supports the assumption that the far exemplars of the high-variability category do not influence classification of the intermediate exemplars, which was used in making predictions for the GCM.

Results

Average results. Participants were very accurate in their training classifications. On average, the mean proportion of correct responses in training was .91. A six-way ANOVA (Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse × Condition × Category) was run to check that none of the counterbalanced factors, or the category structure, affected training performance. There were no significant main effects, largest $F(1, 16) = 2.03$, $p = .17$.

Performance on old training exemplars was also excellent during transfer (see Table 3). A six-way ANOVA (Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse × Condition × Category) was run to examine whether any of the control factors had an effect on performance and to check that performance on old training exemplars was equal for each category. There was a main effect of learning order, $F(1, 16) = 5.84$, $p < .05$, that corresponds to a small (3%) accuracy advantage for the participants learning the 1:2 condition before the 1:2 Expanded condition. Such an increase in accuracy should sharpen a generalization gradient, but it should not lead to an increase in the proportion of responses to one category, which is what is of interest here. There were no other significant main effects, largest $F(1, 16) = 1.84$, $p > .05$. This means no other counterbalanced factor had a significant effect on old training exemplars classification in transfer.

Table 3
Mean Proportion of High-Variability Responses to Old Training Exemplars in Test for Experiment 3

	Condition	
Category	1:2	1:2 Expanded
Low variability	.07 (.01)	.10 (.02)
High variability	.93 (.01)	.93 (.01)

Note. Numbers in parentheses are standard errors of the means.

Each new test exemplar was of equal distance from the nearest exemplar of the low-variability category between the two conditions. (That is, the effect of nearest neighbors was controlled across the two conditions without the adjustment required in Experiment 2.) As in the previous experiment's analysis the responses given to each of the 21 new transfer exemplars were collapsed into seven sets. Figure 7 plots the generalization gradient. A six-way ANOVA (Condition × Stimulus Set × Category Mean and Variance Assignment × Category Label × Condition Order × Rectangle or Ellipse) was run. In both the 1:2 and the 1:2 Expanded conditions, the proportion of high-variability responses to test exemplars increased as the location of the test exemplar moved toward the high-variability category, $F(6, 96) = 277.20$, $p < .05$ (Huynh–Feldt $\varepsilon = 1.00$). There was almost no difference between the proportion of high-variability responses in the 1:2 and 1:2 Expanded conditions, $F(1, 16) = 0.25$, $p > .05$. There was no significant interaction between stimulus and condition, $F(6, 96) = 0.61$, $p > .05$ (Huynh–Feldt $\varepsilon = 0.74$). None of the counterbalanced factors had a significant effect, largest $F(1, 16) = 3.88$, $p > .05$.

Individual participant results. As for Experiment 2, when generalization gradients were calculated for individual participants, they showed that many participants had very different gradients for the two conditions. The results, averaged across participants, did not represent individual performance well. When the distant exemplars of the more variable category were moved to be more extreme points, 8 participants showed an increase in their proportion of high-variability responses to the transfer exemplars, whereas the remaining 24 showed a decrease. Further, for many of these participants the change was larger than would be expected by chance. As before, a chi-square analysis was performed for each participant, with the trial as the unit of analysis. Nineteen participants showed a significant difference between their responding in the two conditions, 4 increasing and 15 decreasing their proportion

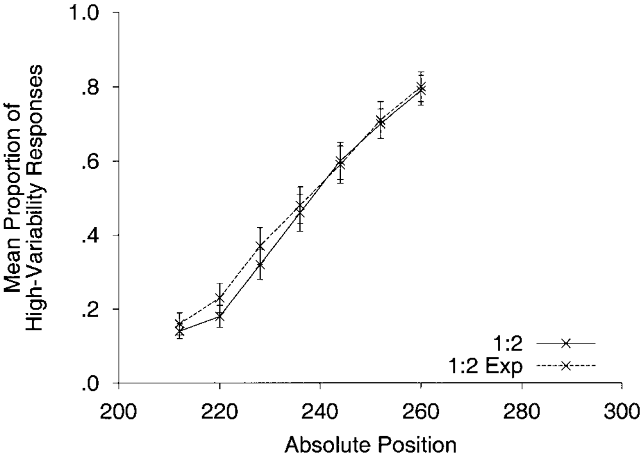

Figure 7. The results of the transfer stage of Experiment 3. Exp = expanded.

of high-variability responses as the difference in variability between the two conditions increased. The probability of obtaining 19 or more significant differences (i.e., $p < .05$) by chance, under the assumption that there is no difference between the proportion of high-variability responses between the two conditions is 3.52×10^{-17}, assuming that the number of significant results is binomially distributed ($n = 32$, $p = .05$).

As previously mentioned, an alternative account of these individual participant data is to postulate random fluctuations in response bias between the 1:2 and 1:2 Expanded conditions. This hypothesis could certainly predict individual differences. Some participants would decrease their bias for the high-variability category in the 1:2 Expanded condition compared to the 1:2 condition. These participants would therefore show a decrease in high-variability-category responses in the 1:2 Expanded condition compared with the 1:2 condition. Similarly, some participants could show the opposite pattern. A key prediction from this random-response-bias hypothesis is that for any participant, the probability of showing either pattern is .5. However only 8 out of 32 participants did show an increase in high-variability responses between the 1:2 and 1:2 Expanded conditions. The probability of 8 or fewer participants showing an increase is .0035, assuming a binomial distribution for the number of participants showing an increase ($n = 32$, $p = .5$). The random-response-bias hypothesis may therefore be rejected. It is possible that there might have been some systematic cause of changes in response bias, which would change the probability of increasing high-variability-category bias between the 1:2 and 1:2 Expanded conditions from a chance level of .5. However, because the order of each condition and the assignment of condition to shapes was counterbalanced across participants, it is not clear what the response bias could vary with, other than the factor of interest—the change in category structure.

Discussion

Moving the distant exemplars of the high-variability category to more distant locations did not alter the generalization gradient obtained from averaged participants' data. This result is consistent with the predictions of the GCM and Normal GRT. However, as in the previous experiment, individual participant data was not well described by the average data. For the majority of participants, moving the distant exemplars had a large effect on their performance on the intermediate exemplars. Both the GCM and Normal GRT are unable to account for this result. Further, significantly more participants than would be expected by chance showed a decrease in the proportion of high-variability responses. Thus the alternative hypothesis raised in the *Discussion* section of Experiment 2—that individual differences are due to random fluctuations between conditions in individual's response biases or perceptual spaces—can be rejected because this hypothesis predicts that increases and decreases in the proportion of high-variability responses should be equally likely. The possibility that these findings might be explained by fluctuations that are nonrandom is not ruled out.

In summary, although average data are consistent with both exemplar and distributional approaches, at the level of individual participants the data for the majority cannot be explained by either approach.

General Discussion

In the experiments presented in this article, we investigated whether categorization performance is based on similarity to stored category exemplars or the likelihood of the data in relation to a probability distribution inferred from the data. Modeling using an exemplar model (the GCM; Nosofsky, 1986) and a distributional model (Normal GRT; Ashby & Townsend, 1986) demonstrated that the two accounts make qualitatively different predictions for the classification of a critical exemplar exactly in-between the nearest exemplars of two categories that differ in variability. The exemplar model predicted classification of the critical exemplar into the more similar, lower-variability category, but the distributional model predicted classification into the more likely, higher-variability category.

Experiment 1 showed that the critical exemplar was classified into the lower variability category most often when stimuli were presented sequentially, consistent with the predictions of the exemplar model. Models of categorization were originally intended to make predictions for sequentially presented stimuli. However, in nonstandard conditions, in which stimuli were presented simultaneously and a hint was given that the two categories differed in variability (manipulations that were intended to increase the salience of the difference in variability), the same critical exemplar was classified into the high-variability category most often, consistent with the predictions of the distributional model. Thus, under some conditions at least, it seems that participants switched from using an exemplar strategy to using a distributional strategy.

Further modeling demonstrated that the exemplar and distributional models make opposite predictions about the effect of increasing the relative variability of the two categories on classification of exemplars intermediate between the two categories. The exemplar model predicted that the probability of classifying an intermediate exemplar into the high-variability category would decrease slightly as the difference in variability increased. At odds with this prediction, the distributional model predicted that the probability of classifying an intermediate exemplar into the high-variability category would increase as the difference in variability increased.

Experiment 2 demonstrated that individual participants' classification of exemplars intermediate between two categories varied greatly as the relative variability of the pair of categories was increased. Some participants showed an increase in high-variability-category responses, consistent with the predictions of Normal GRT, and others showed a decrease, consistent with the predictions of the GCM. The best construal for the GCM and Normal GRT would be that both kinds of mechanism are available to people and they can choose between them. However, this seems to involve the cognitive system in unnecessary duplication, given that the two approaches produce extremely similar answers under almost all circumstances. Moreover, this possibility is eliminated by the results of Experiment 3. Experiment 3 replicated the results of Experiment 2 by using two pairs of categories where both exemplar and distributional models were constrained to predict no change in the proportion of high-variability responses to intermediate exemplars as relative variability was increased. The majority of participants showed a significant change at odds with the predictions of both the GCM and Normal GRT. At the level of data averaged across participants, these differences disappear. That the true form of individual participant data is obscured by averaging further illustrates the dangers of averaging across participants (Ashby et al., 1994; Maddox, 1999).

Exemplar and distributional models can be thought of as lying at opposite ends of a continuum of finite mixture models, where the number of distributions used to represent a category varies from one, as in Normal GRT, to the number of exemplars of that category, as in the GCM and standard GRT (Ashby & Alfonso-Reese, 1995; Rosseel, 1996). (Ashby and Maddox, 1993, and Nosofsky, 1990, also formalize the relationship between exemplar and distributional models.) Also contained in this continuum are back propagation networks with sigmoidal activation functions (Rumelhart, Hinton, & Williams, 1986) and radial basis functions (Moody & Darken, 1989). With small numbers of hidden units (and hence, small numbers of free parameters in relation to the size of the data to be modeled), neural networks are analogous to distributional models because they can learn data only with a particular distributional structure. But if the number of hidden units is large in relation to the amount of data to be learned, then the neural network becomes analogous to an exemplar model in that any data set can be modeled, whatever its structure, simply by learning each piece of data (each exemplar) by rote. The results of Experiments 2 and 3 present a challenge to unitary accounts of this kind that assume that categorization is achieved by a mechanism at some point along the continuum between distributional and exemplar models.

Decision-Bound Models

Decision-bound models of categorization may be adapted to offer a potential account of these results. Decision-bound models include general linear classifiers (e.g., Medin & Schwanenflugel, 1981; Morrison, 1990; Nilsson, 1965; Townsend & Landon, 1983), general quadratic classifiers (e.g., Ashby, 1992; Ashby & Maddox, 1992), and optimal decision rules (e.g., Fukunaga, 1972; Green & Swets, 1966; Noreen, 1981; Townsend & Landon, 1983). Decision-bound models are closely related to Normal GRT, except that participants are assumed to estimate the parameters of the decision bound directly, rather than calculating the bound from the inferred normal distributions used to represent each category.

In the experiments presented in this article, there is a large, empty region between the two categories, where participants have no training data. Therefore, there is a large set of perfect decision bounds that participants could use if they are estimating the bound directly. However, the hypothesis that the individual differences described in Experiment 3 are due to participants choosing a bound at random from the large set of possible bounds in each condition fails. This hypothesis predicts that participants would be as likely to move their decision bound toward the high-variability category in the 1:2 Expanded condition compared with the 1:2 condition as they would be to move it away from the high-variability category. Thus, participants would be as likely to show an increase in high-variability-category responses across conditions as they would be to show a decrease. The finding that the number of participants showing either pattern differs significantly from this chance hypothesis can be used to reject the random-decision-bound hypothesis, just as it was used to reject the random-response-bias hypothesis in Experiments 2 and 3. Thus the selection of the decision bound from the set of possible bounds must be nonrandom. However, decision-bound theory does not provide a candidate selection mechanism. Such a mechanism would also have to account for how the location of this bound might be influenced by knowledge and salience of the differences in variability, as demonstrated in Experiment 1.

Prototype Models

J. D. Smith and Minda (2000) reviewed the categorization literature and found that prototype models (e.g., Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973; Rosch, Simpson, & Miller, 1976) were able to account for performance on novel training exemplars at least as well as exemplar models (although exemplar models out-performed prototype models on old training exemplars). Following this renewed interest in prototype models, the predictions of prototype models for category structures used here are described below.

Prototype models predict classification of exemplars into the category with the nearest mean. Thus, for the critical exemplar in the category structure used in Experiment 1, prototype models predict it should be classified into the low-variability category as the mean of this category is nearest to the critical exemplar. Because the model does not represent variability information, the variability salience manipulations in Experiment 1 should not have had any effect. The category means remain unaltered between the 1:2 condition and the 1:4 condition of Experiment 2, and thus prototype models predict no difference in the (absolute position) generalization gradients between the two conditions. A significant difference was observed, contrary to the predictions of prototype models. For Experiment 3, the motion of the extreme exemplars of the high-variability category to more distant locations (in the 1:2 Expanded condition, compared with the 1:2 condition) will cause the prototype model to predict more high-variability responses to test exemplars in the 1:2 condition than in the 1:2 Expanded condition. In Experiment 3 no significant difference was observed in the average data, and the small numerical difference was in the opposite direction. In summary, prototype models are unable to account for sensitivity to category variability displayed here.

Ashby and Gott (1988)

It is worth noting the relationship between this demonstration that participants are sensitive to the difference in variability of two categories and Ashby and Gott's (1988) Experiment 3. They used a two dimensional category structure with two categories with equal, nonidentity covariance matrices with positive covariance between the two dimensions (illustrated in their Figure 4). The category means differed on a single dimension, and thus the decision bound predicted by a minimum distance (to prototype) classifier is a straight line of equal value on the other dimension between the two categories. The optimal linear decision bound is a diagonal line of positive slope between the two categories. Participants' classification was best described by the optimal linear decision bound, reflecting participants' sensitivity to the correlation of the two dimensions. Thus Ashby and Gott demonstrated that participants were sensitive to within-category covariance. In contrast, the experiments in this article demonstrated that participants were sensitive to the difference in variability between two categories.

Kalish and Kruschke (1997)

Kalish and Kruschke (1997) investigated decision boundaries in a one-dimensional categorization. In their Experiment 1 they used two overlapping uniformly distributed categories of different variance. This structure is therefore similar to that used here in Experiment 1. Although it is perhaps unfair to use Normal GRT to predict performance on Kalish and Kruschke's category structure, as their categories are not normally distributed, the structure does lead to differing predictions for Normal GRT and the GCM. The GCM predicts a two-step generalization gradient, where Normal GRT predicts a one-step function. Kalish and Kruschke found that, of 42 participants, 23 showed a one-step function (i.e., a two-step function did not fit significantly better) and 18 showed a two-step function. These results then provide approximately equal support for either model.

Conclusion

Averaged across participants, under standard conditions of sequential presentation of training exemplars, the data presented here favor an exemplar-similarity based account of classification rather than a distributional account. However, under nonstandard conditions, when training exemplars were presented simultaneously and participants were told that the categories differed in variability, performance switched to that predicted by a distributional account. However, there were large individual differences that neither model could account for when the relative variability of two categories was manipulated. We are beginning to explore an alternative account that differs fundamentally from those discussed here in that the absolute magnitudes of stimulus attributes are assumed to be unavailable, and instead that stimuli are judged relative to one another (Stewart, Brown, & Chater, 2002).

References

Altham, P. M. E. (1979). Detecting relationships between categorical variables over time: A problem of deflating a chi-squared statistic. *Applied Statistics, 28,* 115–125.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 33–53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71.

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

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional-scaling or the similarity-choice model. *Psychological Science, 5,* 144–151.

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

Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 57–105). San Francisco: Freeman.

Feldman, J., & Richards, W. (1998). Mapping the mental space of rectangles. *Perception, 27,* 1191–1202.

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

Fukunaga, K. (1972). *Introduction to statistical pattern recognition.* New York: Academic Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 7,* 418–439.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1362–1377.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12,* 4–34.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Macmillan, N. A., & Ornstein, A. S. (1998). The mean–integrality representation of rectangles. *Perception & Psychophysics, 60,* 250–262.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61,* 354–374.

McLaren, I. P. L. (1997). Categorization and perceptual learning: An analogue of the face inversion effect. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 50*(A), 257–273.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–368.

Monahan, J. S., & Lockhead, G. R. (1977). Identification of integral stimuli. *Journal of Experimental Psychology: General, 106,* 94–110.

Moody, J., & Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation, 1,* 281–294.

Morrison, D. F. (1990). *Multivariate statistical methods.* (3rd ed.). New York: McGraw-Hill.

Nilsson, N. J. (1965). *Learning machines.* New York: McGraw-Hill.

Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology* (pp. 237–279). Providence, RI: American Mathematical Society.

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

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34,* 393–418.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 54*(A), 197–235.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 88,* 304–308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

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.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural base of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 491–502.

Rosseel, Y. (1996). Connectionist models of categorization: A statistical interpretation. *Psychologica Belgica, 36,* 93–112.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, October 9). Learning representations by back-propagating errors. *Nature, 323,* 533–536.

Shepard, R. N. (1980, October 24). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398.

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

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 3–27.

Stewart, N. (2001). *Perceptual categorization.* Unpublished doctoral dissertation, University of Warwick, England.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 3–11.

Tavaré, S., & Altham, P. M. E. (1983). Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika, 70,* 139–144.

Townsend, J. T., & Landon, D. E. (1983). Mathematical models of recognition and confusion in psychology. *Mathematical Social Sciences, 4,* 25–71.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 51*(B), 235–270.

Received June 16, 2000
Revision received June 5, 2001
Accepted March 11, 2002 ■