Memory & Cognition
2005, 33 (2), 289-302

Effects of category diversity on learning, memory, and generalization

ULRIKE HAHN, TODD M. BAILEY, and LUCY B. C. ELVIN
Cardiff University, Cardiff, Wales

In this study, we examined the effect of within-category diversity on people's ability to learn perceptual categories, their inclination to generalize categories to novel items, and their ability to distinguish new items from old. After learning to distinguish a control category from an experimental category that was either clustered or diverse, participants performed a test of category generalization or old–new recognition. Diversity made learning more difficult, increased generalization to novel items outside the range of training items, and made it difficult to distinguish such novel items from familiar ones. Regression analyses using the generalized context model suggested that the results could be explained in terms of similarities between old and new items combined with a rescaling of the similarity space that varied according to the diversity of the training items. Participants who learned the diverse category were less sensitive to psychological distance than were the participants who learned a more clustered category.

This article examines people's ability to learn categories by induction over exemplars, and in particular, how the diversity of training exemplars affects the rate of learning, the pattern of generalization, and the ability to distinguish new exemplars from old ones. Although we focus on perceptual categories, the relevance of diversity to generalization applies widely to people's ability to infer properties of things as in, for example, inductive inference or conceptual reasoning, as well as perceptual classification. In normative terms, diverse evidence gives rise to stronger inductive arguments. This diversity principle has been emphasized in the philosophy of science (for a recent discussion, see, e.g., Wayne, 1995), and considerable experimental work has examined the extent to which it is adhered to in everyday judgments by both adults (e.g., Osherson, Smith, Wilkie, Lopez & Shafir, 1990) and children (for a recent discussion, see, e.g., Heit & Hahn, 2001). With respect to semantic concepts, some categories are cognitively favored over others. An issue that has provoked much discussion is the extent to which the "coherence" of categories relies on similarities and is degraded by diversity (cf. Barsalou, 1983; Corter & Gluck, 1992; Hahn & Ramscar, 2001; Jones, 1983; Murphy & Medin, 1985; Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). However, the relevance of diversity to perceptual categories is uncontroversial. Studies of perceptual categories have shown that even infants are sensitive to category variability (e.g., Mareschal, French, & Quinn, 2000; Quinn, Eimas, & Rosenkrantz, 1993; Younger, 1985). Many studies of adults (reviewed below) have similarly found that diversity influences category processing. The nature of that influence remains unclear.

We can distinguish several fundamental mechanisms by which diversity might influence category processing. First, individual item similarities could produce diversity effects. With respect to generalization, if one has encountered a wide range of examples in the past, a novelty selected at random from a wide range of possibilities is more likely to be similar to something familiar than if one has only a narrow base of highly similar experiences on which to draw. Alternatively, there could be *general* effects of diversity that are independent of individual item similarities. For example, category boundaries might be shifted away from diverse categories, or similarity relations among stimuli might be normalized according to the amount of variability among the stimuli.

Our aim in this article is to distinguish among these different mechanisms and to determine the extent to which variability gives rise to general changes in how categories are processed. To assess the generality of diversity effects across tasks, we examine category learning, generalization, and memory for category members. We also assess the generality of diversity effects across the stimulus space from the outer edges of our categories to their centers and their boundaries.

Category Learning

Several studies have reported that more-variable categories are harder to acquire than less-variable categories (Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Peterson, Meagher, Chait, & Gillie, 1973; Posner, Gold-

L.B.C.E. was supported by the Biotechnology and Biological Sciences Research Council and the Unilever Corporation. We thank Jacky Boivin for valuable advice on the MANOVA and Robert Nosofsky, Dorrit Billman, and several anonymous reviewers for helpful comments on earlier drafts of this article. Correspondence should be addressed to U. Hahn, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, Wales (e-mail: hahnu@cardiff.ac.uk).

 Copyright 2005 Psychonomic Society, Inc.

smith, & Welton, 1967; Posner & Keele, 1968). However, some of these studies are open to alternative interpretations, so the evidence that diversity affects category learning is less overwhelming than it first appears. In the studies by Posner and colleagues, involving random distortions of prototype images, stimulus variability is confounded with recognizability—participants in the low-variability condition, but not the high-variability condition, recognized three out of four category prototypes as familiar shapes (a triangle, an "M," and an "F"). Although the Posner studies are therefore unconvincing as evidence that more-variable categories are harder to learn, two subsequent studies found differences in learnability due to exemplar diversity even when prototype images were random dot patterns rather than familiar images (Homa & Vosburgh, 1976; Peterson et al., 1973).

The only other study demonstrating an effect of diversity on category learning, Fried and Holyoak (1984), used random checkerboard patterns as prototypes and created exemplars of each category by randomly inverting the color of some squares relative to the prototype. Diverse categories (with many squares inverted relative to the prototypes) were harder to learn than less-variable categories (with fewer squares inverted). As Stewart and Chater (2002) point out, the probability that a sizable chunk of a checkerboard pattern remains constant is higher among less-diverse patterns than among those with more variability. Therefore, if discrimination between checkerboard categories relies on chunks common to stimuli in the same category (McLaren, 1997; Palmeri & Nosofsky, 2001), Fried and Holyoak's effect of diversity suggests that it is easier to abstract a few large features from less-diverse patterns than a larger number of small features from more-diverse patterns. A similar interpretation applies to the results discussed above involving distorted dot patterns.

If this explanation is correct—that learning difficulty is determined solely by the number of features required to distinguish the target categories—variability within a fixed number of features should not affect learnability. The categories in our study have been designed to test this prediction. They are distinguished by their locations relative to two obvious, continuous perceptual dimensions; the stimuli are not composed of numerous component parts that could be aggregated into abstract features or invariant chunks.

Generalization to Novel Instances

How does diversity during category learning affect the subsequent classification of novel items? Many studies have shown that generalization is affected by the diversity of category members, but the effect of diversity appears to vary with the location of test items relative to the category prototypes. Far away from a prototype, the diversity of previously seen category members has a positive effect on generalization of that category. This has

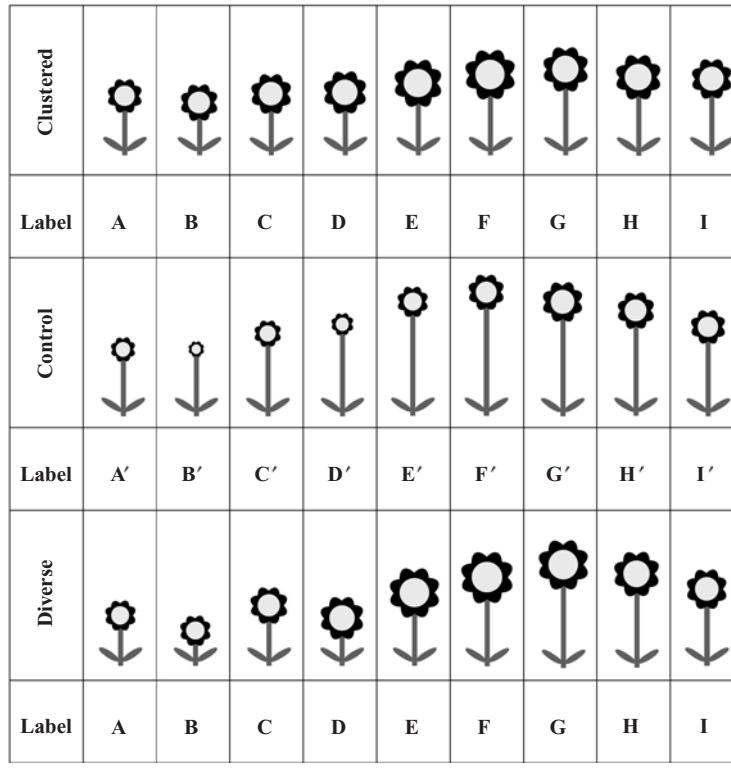

Figure 1. Training stimuli for the clustered and diverse versions of the experimental category and for the control category.

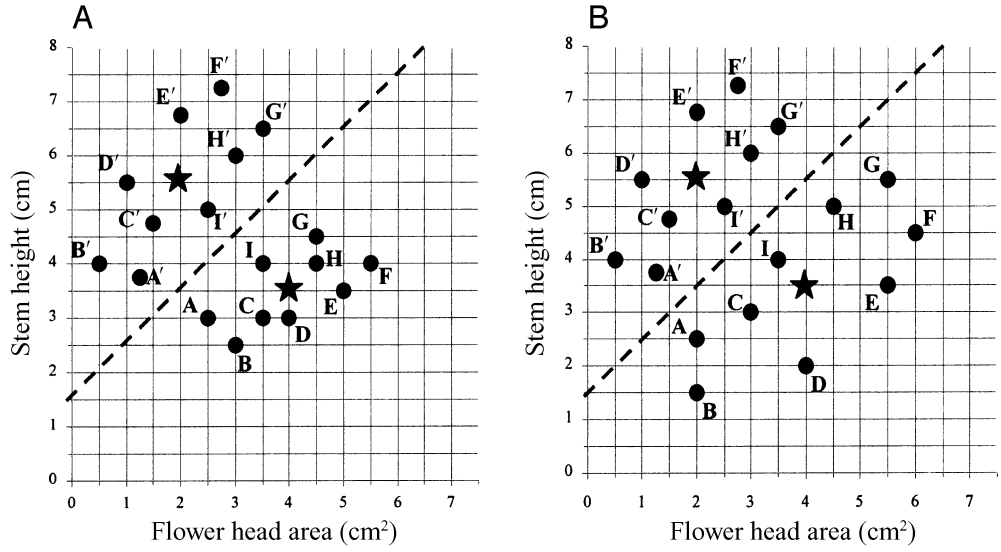

Figure 2. Positions of training stimuli (marked by labeled dots) within the parameter space defined by head area and stem height. One group of participants learned to distinguish the control category from the clustered category X_C (A), and the other group learned to distinguish the same control category from the diverse category X_D (B). Category prototypes (not seen during training) are indicated by stars, and a linear category boundary is shown as a dashed line midway between the prototypes.

been shown both for items near the boundary between categories (Cohen, Nosofsky, & Zaki, 2001; Fried & Holyoak, 1984; Rips, 1989; for examples of boundary items, see Figure 2, items A, I, and G) and for peripheral items on the outer fringes of previously seen exemplars (Flannagan, Fried & Holyoak, 1986; Fried & Holyoak, 1984; Homa & Vosburgh, 1976; Posner & Keele, 1968; for examples of peripheral items, see Figure 2, items D, E, and F). In contrast, category diversity has a negative effect on generalization in the vicinity of the prototype (Flannagan et al., 1986; Fried & Holyoak, 1984; Homa & Vosburgh, 1976).[1] A noteworthy exception to the general pattern was reported by Peterson et al. (1973), who observed negative effects of category diversity on generalization across a wide range of distances from category prototypes. However, the criterion-learning task used by Peterson et al. meant that participants learning higher diversity categories received more training than did those learning lower diversity categories. This difference in amount of training could explain the uni-

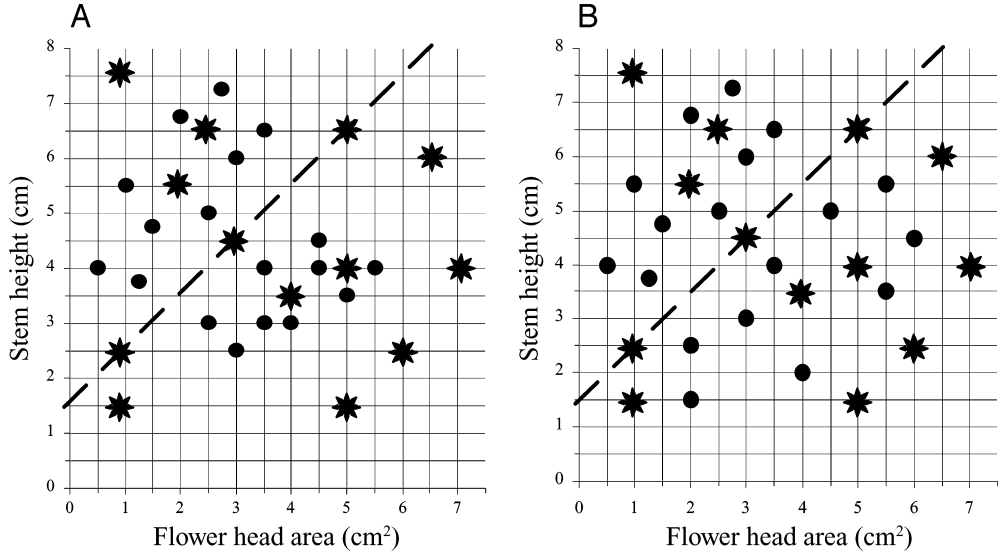

Figure 3. Positions of novel test items (marked by asterisks) and training items (dots) within the head area and stem height parameter space, for participants in the clustered category condition (A) and the diverse category condition (B).

formly negative effects of diversity observed by Peterson et al. (cf. Flannagan et al., 1986).

In summary, the effect of diversity on generalization seems to vary, so that results for boundary items do not necessarily carry over to central items, or to peripheral items, and so on. One possibility is that these various effects simply reflect differences in item similarities. Exemplars of a diverse category are, on average, more distant from the category prototype (see Figure 3), which could account for a negative effect on generalization to a previously unseen prototype. At the same time, these exemplars are closer to the category boundary and also to fixed peripheral items, which could account for a positive effect on generalization. Are there any effects of diversity above and beyond item similarities, and if so, are these effects the same across the stimulus space? Some studies have controlled or factored out item similarities (Cohen et al., 2001; Posner & Keele, 1968; Rips, 1989; Stewart & Chater, 2002; but see the discussion in E. E. Smith & Sloman, 1994, and Nosofsky & Johansen, 2000), and some have examined generalization right across the stimulus space (Flannagan et al., 1986; Fried & Holyoak,1984; Homa & Vosburgh, 1976; Peterson et al., 1973). However, no study has done both. Our study was designed to test for general effects of diversity after item similarities were factored out, while probing generalization across boundary, central, and peripheral items.

We also aimed to distinguish between two basic mechanisms that might alter generalization in response to variability. First, response biases (i.e., a basic tendency to favor one response category over another) might form in the classification decision process in favor of more-variable categories (Ashby & Maddox, 1990; Cohen et al., 2001; see also Stewart & Chater, 2002). Second, similarity comparisons might be rescaled or normalized by the amount of variability observed within the relevant categories, altering the perception of distance in the relevant stimulus space. Thus, if we did observe a general effect of variability, our analyses were designed to distinguish whether the effect was mediated by a response bias or some form of rescaling.

Memory for Instances

How does diversity during category learning affect our ability to distinguish new instances from old? Although some theorists maintain that categorization does not rely on memory for specific exemplars, no one denies that exemplars experienced during category-learning tasks normally form traces in (episodic) memory. Are those memory traces affected by the diversity of items experienced during learning? The mechanisms of response bias and rescaling, discussed above in relation to generalization, may also be relevant to instance memory. Empirically, we are aware of only one study that has examined the effect of diversity in a category-learning task on subsequent recognition memory. Neumann (1977) trained people on a single category, then asked them to rate the familiarity of novel items ("How certain are you that you saw this during training?"). Responses varied depending on the distribution of training items. This result showed that the perceived familiarity of new items was affected by the diversity of items in the training category. However, it is likely that some or all of the effect observed by Neumann reflected item similarities, so it is not clear whether more general effects of diversity were involved.

OVERVIEW OF THE PRESENT STUDY

The goal of our study was to test the effect of category diversity on learning, generalization, and instance memory. Each participant in our study learned a control category, which was the same for all participants, and an experimental category that was either clustered or diverse. This design made it possible, in principle, to detect response bias effects due to category diversity, and also to distinguish between global rescaling across the whole stimulus space and category-specific rescaling around each category according to its own variability. The distance between category prototypes and between the nearest exemplars of the control and experimental categories was the same for all participants, as was the amount of training.

Our study used simple two-dimensional materials: flowers that varied in stem height and head area, as shown in Figure 1. These materials allowed us to manipulate category variability without affecting the meaningfulness of the stimuli (cf. Peterson et al., 1973; Posner & Keele, 1968). Also, with these materials we could plausibly assume that people's category representations would be based on two perceptual dimensions (corresponding closely to stem height and head area), which made it possible to factor out effects of item similarity in regression models. Some reassurance for the justification of this assumption is provided by the fact that the two dimensions are assigned equal weights in best-fit regressions (see Appendixes A and B). Optimal classification accuracy for our stimuli requires equal attention to the dimensions of head area and stem height. If Cohen et al. (2001) are correct that learners distribute attention optimally, the equal weighting we observe lends credence to a close correspondence between psychological and physical dimensions for our stimuli.

Training

In a supervised classification training task, each participant learned the control category (C) and an experimental category (X). All participants had the same number of training trials for all exemplars. This controlled exemplar frequency, which can enhance generalization (e.g., Nosofsky, 1988), but did not guarantee the same level of category acquisition across both categories or both groups as training to criterion performance would. We manipulated the dispersion of training items around the category X prototype, so that half the participants learned a clustered category (X_C) and half learned a diverse category (X_D), as shown in Figure 2.

If it is generally more difficult to learn highly variable categories than to learn less variable ones, even when the number of distinguishing features is held constant, participants learning X_D should make more errors, require more training before achieving criterion performance, and give slower responses than participants learning X_C (though previous studies have not assessed whether category diversity affects response speed during learning). Moreover, these effects might be either global, affecting categories *C* and *X* equally, or category specific, affecting only category *X*.

Generalization and Recognition

After the training task, each participant completed either a generalization or old–new recognition task involving the novel items shown in Figure 3. These include the previously unseen category prototypes, internal items within the region spanned by the training items, boundary items that were equidistant from the two category prototypes (on the dashed lines in Figures 2 and 3), and peripheral items beyond the outermost training items. In addition to the novel items, which were the same for all participants, the generalization task included some of the training items to reinforce category diversity during the task. The recognition task included all of the training items in addition to the novel ones.

Participants in the generalization task classified items as members of category *C*, category *X*, or *neither* (as in Peterson et al., 1973). The *neither* response provides participants with an alternative to random guessing for peripheral items that may bear little resemblance to either training category. The availability of a *neither* response elicits up to 60% more information compared with a two-alternative forced-choice task and reveals the structure of each category on the far side as well as near the boundary between them. The ratio of *C* versus *X* responses should be unaffected by the availability of a *neither* response (Luce, 1959). To see this, suppose that without *neither* responses a particular peripheral item is classified in category *C* rather than *X* 9 out of 10 times. With *neither* responses, suppose that 80% of the category *C* responses for this item change to *neither*, along with the same fraction of category *X* responses. The ratio between *C* and *X* responses remains unchanged.

On the basis of item similarities and consistent with previous results for items at various locations within the stimulus space (e.g., Fried & Holyoak, 1984, and other works discussed above), we predicted that, on novel items far from the prototype, the diverse category, X_D, would generalize more than the clustered category, X_C, producing more category *X* responses to boundary items and to peripheral items beyond the category *X* training exemplars. For the prototype of the experimental category, we predicted that the diverse category would generalize less, producing fewer category *X* responses to this item than the clustered category would. Following Fried and Holyoak, we also predicted that people who learned the diverse category would give fewer *neither* responses overall to novel items than would people who learned the clustered category.

In the recognition task, participants classified each item as *old* if they thought they had seen it during the training task, or else *new*. There are virtually no relevant previous studies from which to derive detailed predictions, but if diversity of training items affects recognition, people who learned the diverse category should produce more false alarms to novel items than people who learned the clustered category.

To factor out differences in item similarities between the two groups, we fit regression models that predicted generalization and recognition responses to all test items (old and new) on the basis of their similarities to the training exemplars. Analyses tested whether category diversity affected response biases and whether there were either global or category-specific differences in distance scaling. Our regression analyses were based on the generalized context model (GCM; Nosofsky 1984, 1986; see also Medin & Schaffer, 1978). However, our interest was not to test a particular theory of categorization or recognition, but to test whether item similarities alone were sufficient to explain any differences we observed between groups. The GCM has a well-documented capacity to model exemplar effects for perceptual classification stimuli such as ours, and it provides a powerful analytical tool to test for effects above and beyond the influence of item similarities. Details of our adaptations of the GCM are given in Appendixes A and B.

METHOD

Participants

The participants were 73 undergraduate psychology students who received course credit for participating. All but six were female. The participants were randomly assigned to one of two training conditions, with 36 participants 18–24 years old ($M = 19.44$, $SD = 1.27$) in the clustered category condition and 37 participants 18–31 years old ($M = 19.54$, $SD = 2.19$) in the diverse condition. No color-blind participants were recruited. After the training task, 41 participants did the generalization task (20 and 21 participants from the clustered and diverse conditions, respectively). The other 32 participants did the recognition task (16 each from the two training conditions).

Stimuli

The training task used three sets of nine flowers, depicted in Figure 1, comprising exemplars of the control category, *C*, the clustered category, X_C, and the diverse category, X_D. There were three variants of each flower with different colors in the flower head. Figure 2 shows the locations of the training exemplars within the two-dimensional space of stem height and head area. Categories X_C and X_D had the same prototype, that is, the same stem height and head area averaged across training exemplars. Apart from exemplar *I*, which is the same in X_C and X_D, all other exemplars were closer to the prototype in X_C than in X_D. Exemplars of the control category corresponded to pairwise average coordinates of X_C and X_D exemplars, reflected across the diagonal category boundary shown as a dashed line in Figure 2. Category *C* was therefore a mirror image average of X_C and X_D, with an intermediate level of diversity. Neither stem height nor head area alone were sufficient to distinguish category *C* from X_C or X_D—participants had to use both stimulus dimensions to accurately classify members of the two categories.

Thirteen novel flowers, for the generalization and recognition tasks, were located as in Figure 3. We included more peripheral items for category *X* in anticipation of the diversity manipulation affecting it more than category *C*. The novel items for category *C* occupied mirror image coordinates with respect to corresponding items for category *X*. All 13 novel flowers were used in the generalization task, along with training exemplars *A′*, *D′*, *F′*, *H′*, and *I′* from category *C* and the corresponding exemplars from either X_C or X_D, as appropriate. The recognition task used the novel flowers, plus the exemplars of category *C* and either X_C or X_D, as appropriate.

Three different flower head color schemes were used to create a task-irrelevant dimension of variation, with three variants of each flower. During training, each participant saw each flower the same number of times in each color scheme. In the generalization and recognition tasks, each flower occurred once in each color scheme. In addition to reducing ceiling effects during training, the color variations allowed us to probe recognition memory for each flower three times (once in each color scheme), without actually presenting any specific stimulus more than once.

Procedure

The participants were tested individually in a quiet room. First, the experimenter read instructions for the learning task aloud. The participants were encouraged to respond quickly, but accurately. A computer controlled by a SuperLab script presented stimuli on a display screen and recorded the participants' keypresses and latencies. During training, the participants classified flowers into a "gold" category (category *C*) and a "silver" category (category *X*) by pressing keys labeled with an appropriately colored star. After each response, the flower image was replaced with feedback saying either CORRECT or INCORRECT for 500 msec before the next flower was presented. There were 15 blocks of training trials, with each of the 18 training flowers presented once per block in a different random order. Random orders were matched across participants in the two training conditions. Colors were randomized, but each flower appeared five times in each color scheme. The participants had a short break after every three blocks.

After training, each participant performed either the generalization or old–new recognition task. Both tasks began with instructions read aloud by the experimenter. In the generalization task, the participants classified flowers into the categories "gold," "silver," or "neither" (a blue sticker marked the key for a *neither* response). There were three blocks of generalization trials, with 13 novel flowers and 10 old flowers in each. Colors were randomized, with each flower appearing once in each color scheme. In the old–new recognition task, the participants decided whether each flower had appeared during the training task, and responded by pressing keys labeled *old* or *new*. There were three blocks of recognition trials, with 13 novel and 18 old flowers in each. Again, colors were randomized, with each flower appearing once in each color scheme. The participants received no feedback during either the generalization or recognition tasks. Random orders were matched across participants in the clustered and diverse training conditions. After each classification or old–new response, the test flower was replaced with the question "How confident are you?" and the participants responded on a scale from 1 (*low confidence*) to 9 (*high confidence*). The participants had the opportunity for a short break between each block.

RESULTS

Learning

The first block of trials was excluded from analysis, because it would primarily reflect guessing. Participants learning the diverse category generally made more errors than those learning the clustered category, as shown in Figure 4. Inferential tests were based on three measures of difficulty computed for each participant: the first training block completed with no errors, the total number of errors made, and the mean reaction time (RT) for correct responses. Here and throughout, two-tailed tests are reported, and significance is evaluated at an alpha level of .05.

In computing the first error-free training block (from 2–15), two participants learning the clustered category and 10 learning the diverse one did not achieve an error-free block. These participants were assigned scores of 16. Scores for the first error-free block did not satisfy parametric assumptions, so this variable was analyzed separately from errors and RTs. Participants learning the clustered category achieved their first error-free block earlier in training [median = 4.0, IQR (interquartile range) = 6.75] than did participants learning the diverse

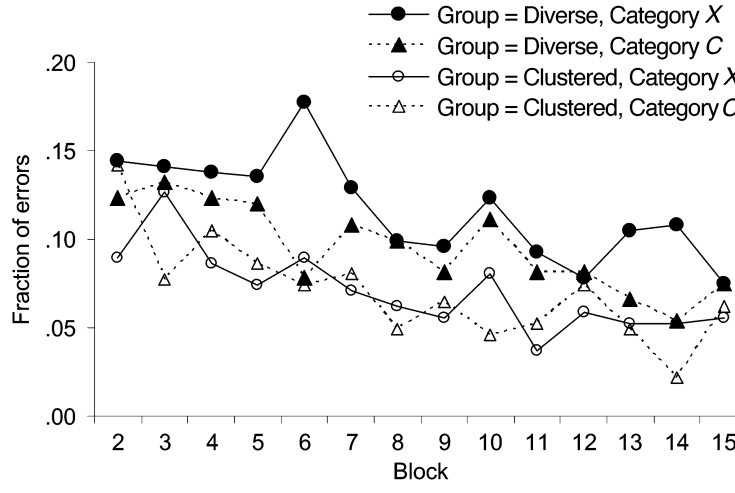

Figure 4. Fraction of error responses during category learning, as a function of training group (clustered or diverse), category (*C* or *X*), and block of training.

Table 1
Summary Measures of Training Difficulty, Including Error Rates and Reaction Times (RTs, in Milliseconds) for Correct Responses, as a Function of Training Group (Clustered or Diverse) and Test Category (*C* or *X*)

	Error Rate (%)		RTs	
Training	Category *C*	Category *X*	Category *C*	Category *X*
Clustered	5.5	6.1	595	601
Diverse	8.3	10.3	685	705

category (median = 9.0, IQR = 12.0). This difference was significant in a Mann–Whitney test ($U = 446.0$, $n_1 = 36$, $n_2 = 37$, $p = .014$).

The total numbers of errors were computed separately for the control and experimental categories, and a square root transform was applied to the totals for each participant to reduce skewness, kurtosis, and heterogeneity of variance. The mean root errors were 2.62 and 2.77 for the clustered condition categories *C* and *X*, respectively, and 3.24 and 3.60 for the diverse condition ($SE = 0.238$, 0.189, 0.206, and 0.226, respectively). These mean values are detransformed and expressed as error rates in Table 1. RTs for correct responses were log transformed. Three outlying RTs (out of 16,765), less than 200 msec or greater than 5,000 msec, were excluded from analysis. The mean log RTs were 2.775 and 2.779 for the clustered condition categories *C* and *X*, and 2.836 and 2.848 for the diverse condition ($SE = 0.016$, 0.016, 0.014, and 0.015, respectively). These mean values are detransformed and shown in Table 1. A two-way multivariate analysis of variance (MANOVA) was performed on root errors and log reaction times, with training diversity (clustered or diverse) as a between-subjects factor and stimulus category (control or experimental) as a within-subjects factor.[2] Here and throughout, there were no multivariate outliers within levels of training diversity ($p < .001$). This means that the other statistics from these MANOVAs are not unduly influenced by a few unrepresentative data points.

In combination, errors and RTs were affected by category diversity [$F(2,70) = 7.67$, $p = .001$]. People learning the clustered category made fewer errors and gave faster responses than did people learning the diverse category. There was no significant difference between the control and experimental categories [$F(2,70) = 1.62$, $p = .21$], nor an interaction between diversity condition and category [$F(2,70) < 1$]. Thus, the diversity of category *X* affected responses to both categories *C* and *X* equally.

Errors and RTs made independent contributions to the combined differences between training groups, as indicated by a unique effect for each variable after factoring out the other in Roy–Bargmann stepdown analyses [Fs$(1,70) > 5.24$, ps $< .025$]. In univariate ANOVAs, the main effect of category diversity was significant both for errors [$F(1,71) = 7.78$, $p = .007$] and for RTs [$F(1,71) = 9.51$, $p = .003$].

Generalization

Average response percentages for various types of test item in the category generalization task are shown in Table 2. Asterisks identify those entries that are relevant to the predicted diversity effects we outlined above. To limit the family-wise error rate, we tested only a few key contrasts from this table, and followed up with regression analyses encompassing the whole of the data. Confidence ratings for the generalization and recognition tasks have not yet been analyzed and are not reported.

Table 2
Distribution of Generalization Responses, Showing Percentage of Each Response Alternative as a Function of Stimulus Region and Type of Novel Item

			Response					
			Control		Experimental		Neither	
Stimulus Region	Item Type	Training	%	*SD*	%	*SD*	%	*SD*
Boundary	Boundary	Clustered	54	5	18*	4*	28	4
		Diverse	53	5	23*	4*	24	4
Control	Prototype	Clustered	95	3	2	2	3	2
		Diverse	95	3	0	0	5	3
	Inside	Clustered	97	2	0	0	3	2
		Diverse	95	3	2	2	3	2
	Peripheral	Clustered	90	6	5	5	5	4
		Diverse	87	7	0	0	13	7
Experimental	Prototype	Clustered	2	2	88*	4*	10	4
		Diverse	2	2	76*	7*	22	7
	Inside	Clustered	0	0	92	3	8	3
		Diverse	2	2	79	7	19	7
	Peripheral	Clustered	6	2	50*	4*	44	4
		Diverse	8	2	66*	4*	26	3
Overall		Clustered	37	3	38	3	26*	2*
		Diverse	37	3	43	3	21*	2*

*Entries relevant to predicted diversity effects outlined in text.

Analysis of response fractions. Our initial analysis of generalization examined four measures: the fraction of *neither* responses across all novel items (neither), the fraction of category *X* responses for novel boundary items (boundary*X*), the fraction of category *X* responses for novel items on the periphery of category *X* (periph*X*), and the fraction of category *X* responses for the unseen category *X* prototype (proto*X*). The proto*X* scores were near ceiling and could not be analyzed parametrically. Although proto*X* scores were slightly lower for participants trained on the diverse category compared with the clustered category, this difference between groups was not significant [Mann–Whitney (corrected for ties) $Z = 1.16$, $n_1 = 20$, $n_2 = 21$, $p = .25$].

A one-way MANOVA was performed on the variables neither, boundary*X*, and periph*X*, with training diversity (clustered or diverse) as a between-subjects factor. In combination, these variables were affected by training diversity [$F(3,37) = 3.19$, $p = .035$]. People trained on the diverse category gave fewer *neither* responses overall, more category *X* responses for boundary items, and more category *X* responses for items on the periphery of category *X*, as predicted. Roy–Bargmann stepdown analysis (details not reported here) indicated that the significance of the overall effect was due primarily to the periph*X* variable—people trained on the diverse category gave fewer *neither* responses to items on the periphery of category *X* and classified them as members of category *X* instead.

Regression models. In our stimuli, similarities between test and training items necessarily varied as a function of training diversity, so it is possible that the effect of diversity on periph*X*, for example, reflects the fact that peripheral items were closer to some of the diverse training exemplars than to the corresponding clustered exemplars. To factor out the effect of item similarities and to test for a general effect of variability, we fit regression models as described in Appendix A to predict each participant's responses to the full set of test items (both old and new). The predicted probability of a particular response was a function of the similarity between each test item and the training exemplars of categories *C* and *X*. The full model included three parameters of interest: (1) a distance scaling parameter, *s*, that determined the sensitivity of response probabilities to a given change in head area or stem height, (2) a category *C* relative response bias, $\text{bias}_{C|CX}$, that determined the bias (on a scale from 0 to 1) to classify something as category *C* rather than *X*, and (3) a *neither* response bias, $\text{bias}_{\text{neither}}$, that determined the bias to say that something belonged to neither category *C* nor *X*. Changes in *s* and $\text{bias}_{\text{neither}}$ can have superficially similar effects on predicted response probabilities, but a change in $\text{bias}_{\text{neither}}$ will not affect the ratio of *C* to *X* responses, whereas a change in sensitivity will. The two parameters consequently have distinguishable effects on regression fits.

Goodness of fit statistics for models with and without separate $\text{bias}_{C|CX}$ parameters for each participant indicated that it did not vary significantly within training groups [$\chi^2(39) = 41$, $p = .37$], nor between training groups [$\chi^2(1) = 0.21$, $p = .65$]. Subsequent analyses fixed this parameter at the overall median value, which favored category *C* over *X* by a ratio of more than 2:1. This level of bias means that an item that was equally similar to both categories would be classified as a member of category *C* twice as often as category *X*. This strong bias may reflect participants' expectation that the two categories would occur equally often, whereas there was a preponderance of novel items on the category *X* side of the stimulus space.[3]

With $\text{bias}_{C|CX}$ held at the overall median value, we determined best-fit values of *s* and $\text{bias}_{\text{neither}}$ for each participant. Figure 5 shows response contours for the two training groups, based on median values of *s* and $\text{bias}_{\text{neither}}$. As the 70% contours show, the inclusion of a *neither* response allowed us to model category structure on the far side of each category as well as on the side near the boundary. The median values of $\text{bias}_{\text{neither}}$ were 0.0297 and 0.0310 for the clustered and diverse training groups, respectively (IQRs = 0.0446 and 0.0549). This difference was not significant in a Mann–Whitney test ($U = 175$, $n_1 = 20$, $n_2 = 21$, $p = .36$).

The median values of *s* were 2.72 and 2.26 for the clustered and diverse groups (IQRs = 1.40 and 0.55).

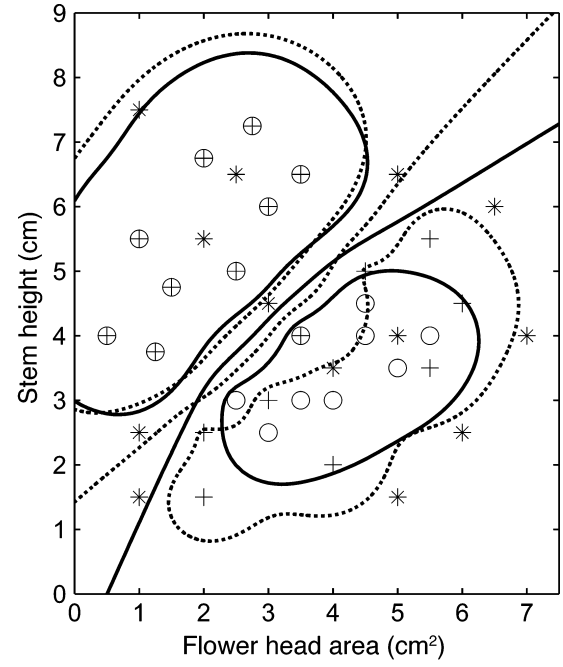

Figure 5. Categorization training and test items and predicted response contours. Markers show training items for clustered (○) or diverse (+) training groups and novel test items (✳). Solid lines are for the clustered training group, showing the equiprobability category boundary and the 70% categorization contours for each category. Dotted lines show the same information for the diverse training group.

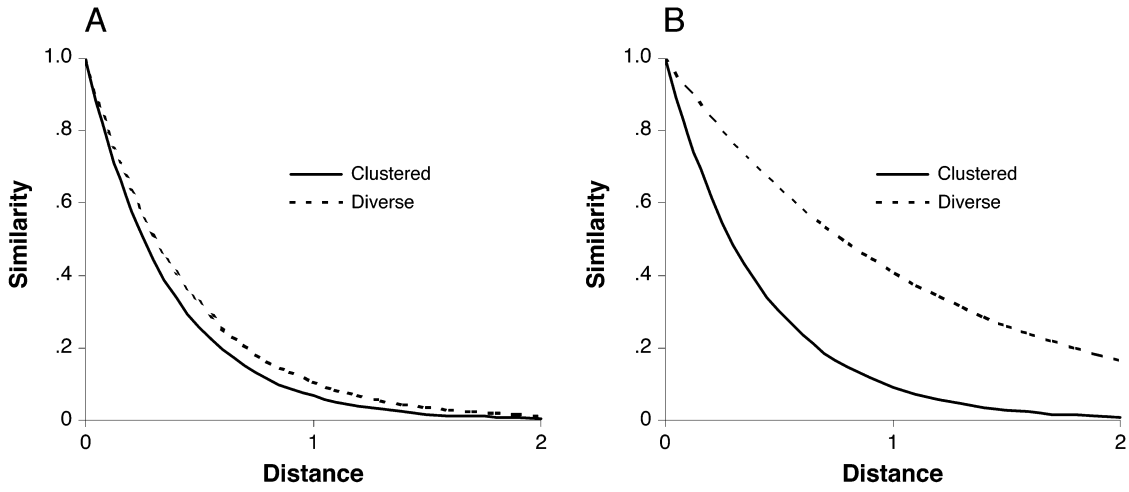

Figure 6. Generalized context model similarity as a function of training group (clustered or diverse) and distance from an exemplar. Derived from median best-fit values of *s* for each training group, when tested on generalization (A) and recognition (B).

Participants who learned the diverse category had significantly lower sensitivity to distance than did those who learned the clustered category ($U = 131$, $n_1 = 20$, $n_2 = 21$, $p = .039$). The effect of this difference in s values is illustrated in Figure 6A, which shows the influence of each training item on the classification of a test item as a function of the distance between the two. The rescaled similarity of an exemplar at distance 0.4 (where the two curves are furthest apart) is 20% greater for participants who learned the diverse category than for those who learned the clustered category. Distance 0.4 corresponds to a circle around a training item on Figure 2 or 3 extending not quite one grid unit in each direction.

Did the effect of training diversity affect distance scaling for categories C and X equally? The GCM ordinarily assumes that the same scaling applies throughout the stimulus space, but Nosofsky and Johansen (2000) have suggested that category-specific scaling could explain the apparent bias toward diverse categories reported in Rips' (1989) classic pizza-or-coin study. We fit an enhanced regression model that had separate s parameters for categories C and X (for each participant) and found no significant difference in goodness of fit between the enhanced model and the original [$\chi^2(41) = 37, p = .66$]. Our results therefore suggest global rescaling across the whole stimulus space and do not indicate category-specific rescaling.

Recognition

The proportion of incorrect responses for various types of test item in the old–new recognition task are summarized in Table 3. Miss rates for training exemplars ranged from 22% to 33%, whereas false alarm rates for most new items were over 50%. The two training groups differed markedly on their responses to peripheral items beyond the outer edge of category X exemplars. People trained on the clustered category mistook these for old items on only 22% of trials, whereas those trained on the diverse category mistook them on 50% of trials.

Analysis of false alarms. Due to the very limited amount of relevant past research, the predictions (outlined above) for the recognition task were very general, so we conducted only a limited set of direct inferential tests. Our analysis pooled novel items into three groups and examined rates of false alarms (FA) for novel items on the category C side of the stimulus space (FA_C), the category X side (FA_X), and on the boundary between categories (FA_Bndy). A one-way MANOVA was performed on the three false alarm scores, with training diversity (clustered or diverse) as a between-subjects factor. In combination, the false alarm scores were affected by training diversity [$F(3,28) = 5.48$, $p = .004$]. Roy–Bargmann stepdown analysis indicated that the significance of the overall effect was due primarily to FA_X; par-

Table 3
Error Rates in Old–New Recognition Task by Training Diversity Showing Percentage for Training Exemplars (Old) and Novel Test Items (New), as a Function of Stimulus Region and Type of Novel Item

			Training			
			Clustered		Diverse	
	Region	Type	%	*SD*	%	*SD*
Old	Control		25	29	33	30
	Experimental		22	24	29	28
New	Boundary	Boundary	58	37	65	35
	Control	Prototype	83	24	79	24
		Inside	77	29	58	26
		Peripheral	58	35	46	34
	Experimental	Prototype	83	24	77	23
		Inside	69	26	77	26
		Peripheral	22	30	50	36

ticipants trained on the clustered category made fewer false alarms on the category *X* side of the boundary (adjusted for FA_*C* and FA_Bndy, mean FA_*X* = 37.5%, *SE* = 4.47%) than did participants trained on the diverse category (adjusted mean FA_*X* = 58.4%, *SE* = 4.47%). Inspection of false alarm rates for individual items (Table 3) suggested that the effect was largely attributable to peripheral items.

Regression models. The effect of diversity on FA_*X* could be due to item similarities, since the peripheral items on the category *X* side were closer to the nearest exemplars of the diverse category than to those of the clustered one. To factor out item similarities, we fit regression models as described in Appendix B to predict each participant's responses to the full set of test items. The predicted probability of an *old* response was a function of the similarity between each test item and the training exemplars of both categories *C* and *X*. There were two free parameters of interest, including the distance scaling parameter, *s*, as discussed above, and a response bias, B_{new}, that determined the bias (on a scale from 0 to 1) to say that something was *new* rather than *old*. Figure 7 shows response contours for the two training groups, based on median values of *s* and B_{new}. The response probability distribution was generally broader and less peaked in the diverse condition than in the clustered condition.

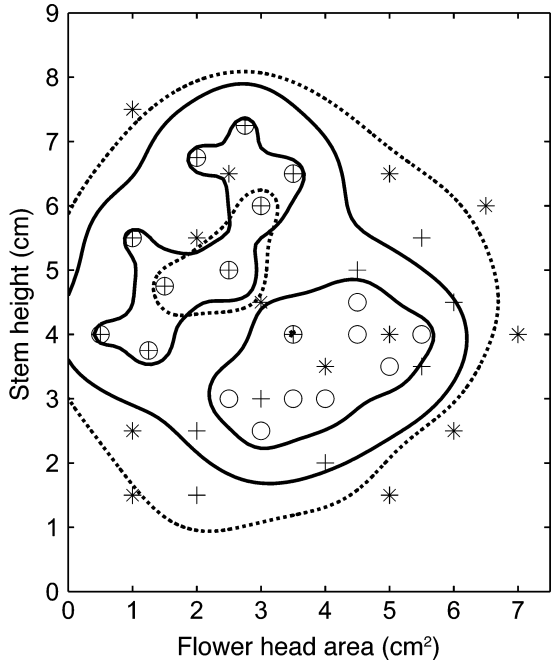

Figure 7. Recognition test items and predicted response contours. Markers show training items for clustered (○) or diverse (+) training groups and novel test items (✳). Solid lines are for the clustered training condition, showing 50% (outer) and 70% (inner) recognition contours. Dotted lines show the same information for the diverse training group.

The median values of B_{new} were .23 and .33 for the clustered and diverse training groups, respectively (IQRs = .26 and .17). However, the difference was not statistically significant ($U = 81$, $n_1 = 16$, $n_2 = 16$, $p = .080$). Thus, as a nonsignificant trend, participants who learned the clustered category were somewhat less inclined to give *new* responses (equivalently, they required somewhat less evidence for an *old* response) than were participants who learned the diverse category.

The median values of *s* were 2.40 and 1.17 for the clustered and diverse training groups, respectively (IQRs = 1.48 and 1.40). Participants who learned the diverse category had significantly lower sensitivity to distance than did those who learned the clustered category ($U = 33$, $n_1 = 16$, $n_2 = 16$, $p < .001$). This difference in sensitivity echoes that observed in the generalization study. The effect of the sensitivity difference on recognition is illustrated in Figure 6B. The difference in exemplar influence is greatest at distance 0.65 (about 1 1/3 grid units in Figures 2 and 3), where the influence is 168% greater for participants who learned the diverse category than for those who learned the clustered category.

Once more, there was no significant improvement in fit for a model incorporating separate scaling parameters for categories *C* and *X* compared with the simpler model with a single global scaling parameter [$\chi^2(32) = 23.4$, $p = .86$].

GENERAL DISCUSSION

Our experiments suggest that the diversity of perceptual categories affects learning, generalization, and item recognition by altering the scale of similarities as well as the positions of exemplars within the stimulus space. People learning a diverse category made more errors and gave slower responses than did people learning a clustered category. After having the same amount of category training, people who learned a diverse category were more likely than those who learned a clustered category to accept distant peripheral items as members of the category. And people who learned a diverse category were more likely to wrongly think that they had previously seen a distant peripheral item than were people who learned a clustered category. The effects of diversity on generalization and recognition were generally consistent with responding on the basis of item similarities. However, regression modeling with the GCM revealed a further effect. Participants' perception of the stimuli was altered, so that people who learned a diverse category required a greater physical difference to perceive the same psychological difference between stimuli. This perceptual rescaling effect led to lower levels of accuracy in the training task, wider generalization, and poorer item recognition, over and above effects that could be attributed to item similarities.

This study sought to distinguish different mechanisms that might produce diversity effects, and the use of three different, but related, tasks makes the present study particularly informative. There are numerous accounts of

category learning in the literature that could explain the reduced *learning* accuracy we observed in the diverse condition, but which do not simultaneously explain the effects we also observed in generalization and recognition. For example, Fried and Holyoak's (1984) category density model represents categories by their means (prototypes) and variances, estimated from a sample of exemplars. Diverse categories are learned more slowly than are clustered ones, because estimates of the means and variances converge more slowly when the stimuli are less uniform. However, because the category density model is concerned only with categorization, it has nothing at all to say about the diversity effects we observed in old–new recognition. And although the category density model predicts that diversity will affect generalization, the pattern of responses in our data are predicted much better by exemplar similarities than by similarities to the category prototypes (cf. J. D. Smith & Minda, 2002; see Appendixes A and B for details).

Diversity effects could plausibly arise from item similarities, shifts in category boundaries (as reflected in categorization biases), or rescaling of similarity relations. A combination of methodological factors contributed to our ability to distinguish these different mechanisms. Regression models allowed us to separate out item similarities, to simultaneously consider response probabilities across a large set of test items (including old and new items, peripheral and internal as well as boundary items) and, in the generalization study, to simultaneously consider the distribution of responses across all three choices. We would not have detected the global rescaling effect if we had examined just a few critical items on the boundary between categories, because the effect of the diversity manipulation on individual items was quite small after item similarities were factored out. In the generalization task, the *neither* response increased the information obtained on each trial (just as a four-choice multiple choice test is more informative than the same number of true/false questions), and made it possible to detect changes in category structure on the far sides of the categories.

Most important, the *neither* response made it possible, in principle, to distinguish between changes in category biases and rescaling of similarities. We found no evidence that similarity relations were category specific, with different scaling factors for different regions of stimulus space according to the variability of categories in those regions. Our data also showed no evidence that response biases varied as a function of category variability. Rather, there was a global change in sensitivity to psychological distance in response to increased variability. This finding is consistent with the effect of stimulus range observed by Braida and Durlach (1972) in studies of sound intensity perception. Our results show that the effect is not confined to unidimensional stimuli and also demonstrate that it affects standard category learning and generalization tasks, as well as item recognition.

With regard specifically to generalization, our modeling demonstrates that exemplar similarities can explain the seemingly different results observed in previous studies for central, peripheral, and boundary items. Only a few previous studies have controlled exemplar similarities in their tests of diversity effects on generalization (Cohen et al., 2001; Rips, 1989; Stewart & Chater, 2002). These studies focused on boundary items, between a clustered category and a diverse category. Rips' pizza or coin study found that judgments of category membership differed from judgments of category similarity, and in particular, that categorization of boundary items was biased in favor of diverse categories rather than clustered categories. Rips concluded that because judgments for the categorization task differed from those for the similarity task, categorization could not be based on similarity. Nosofsky and colleagues have suggested alternative interpretations within the framework of the GCM (Cohen et al., 2001; Nosofsky & Johansen, 2000). Moreover, our own results and those of numerous other studies (e.g., Shin & Nosofsky, 1992; see also works cited in Nosofsky & Johansen, 2000) show that similarity scaling can vary markedly between different tasks. This factor alone could explain the differences that Rips observed between judgments of category membership and category similarity.

The same confound applies to Cohen et al.'s (2001) Experiment 1, which contrasted performance on identification and categorization tasks. On the basis of a single test item, Cohen et al. concluded that high-diversity categories exert a "pull" on category judgments. However, their results could equally be explained by global effects of the sort observed in the present study—that is, by similarity scaling throughout the stimulus space, adjusted according to the particular task, as well as the overall diversity of the stimulus items. In contrast, Cohen et al.'s Experiment 2 provides more compelling evidence for a category-specific effect of diversity on categorization. However, because their analysis focused on a single boundary item, there is no way to tell whether the category-specific effect involved response biases or category-specific rescaling of similarities. These different mechanisms can only be distinguished by examining responses across a wider range of stimuli.

Finally, our finding that category diversity affects global similarity relations has implications for the interpretation of Stewart and Chater's (2002) results. Their analyses of individual data indicated that, for most participants, boundary items were less likely to be classified as members of an experimental category with high diversity than with low diversity. On the basis of these results, Stewart and Chater suggest that response biases might be sensitive to category diversity. Alternatively, their results could reflect the same kind of global rescaling of similarities that we observed in the present study.

Although our results point to a global effect of diversity that affects the whole stimulus space in learning, generalization, and recognition tasks, they do not identify what aspect of diversity is relevant (nor have previous studies of diversity effects in categorization). Participants might be sensitive to the total range of stimuli (Braida & Durlach, 1972), or they might be sensitive to

average absolute deviations from the grand mean, or average variance about category means, or some other specific measure of variability. Only detailed comparisons across various stimulus distributions will eventually distinguish among these possibilities.

REFERENCES

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 598-612.

Barsalou, L. W. (1983). Ad-hoc categories. *Memory & Cognition*, **11**, 211-227.

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.

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

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, **111**, 291-303.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 241-256.

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.

Hahn, U., & Ramscar, M. C. A. (2001) Mere similarity? In, U. Hahn, and M. C. A. Ramscar (Eds.), *Similarity and categorization* (pp. 257-272). Oxford: Oxford University Press.

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology*, **43**, 243-273.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 322-330.

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, **94**, 423-428.

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

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, **36**, 635-645.

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

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

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316.

Neumann, P. G. (1977). Visual prototype formation with discontinuous representation of dimensions of variability. *Memory & Cognition*, **5**, 187-197.

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 54-65.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, **7**, 375-402.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, **97**, 185-200.

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*, **54A**, 197-235.

Peterson, M. J., Meagher, R. B., Jr., Chait, H., & Gillie, S. (1973). The abstraction and generalization of dot patterns. *Cognitive Psychology*, **4**, 378-398.

Posner, M. I., Goldsmith, R., & Welton, K. E., Jr. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, **73**, 28-38.

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

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, **22**, 463-475.

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

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.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 800-811.

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

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Wayne, A. (1995). Bayesianism and diverse evidence. *Philosophy of Science*, **62**, 111-121.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, **56**, 1574-1583.

NOTES

1. These studies all involved nonnormal distributions of exemplars, which were generally sampled from donut-shaped regions of stimulus space. The effect of changes in the variance of normally distributed training exemplars has received little, if any, experimental attention.

2. The MANOVA computes an optimal linear combination of several dependent variables (DV), and computes an ANOVA on the resulting composite DV. If the original DVs are correlated (positively or negatively), the MANOVA protects against the inflated Type I error rate produced by separate tests of the DVs (see Tabachnick & Fidell, 1996).

3. Thanks to Dorrit Billman for suggesting this explanation.

APPENDIX A
Details of Modeling for Generalization Data

According to the generalized context model (GCM), classification decisions are based on similarity comparisons between a test item and individual exemplars of each category stored in memory. In our generalization task, participants judged whether item *i* was a member of category *C*, category *X*, or neither. The GCM predicts the probability of a category *C* response on the basis of the total weighted similarity of the test item to exemplars of category *C*, divided by the weighted similarity of the item to exemplars of both categories, plus a constant response threshold for *neither* responses:

$$p(\text{response}_C \mid i) = \frac{\text{bias}_C \cdot \sum_{j \in C} e^{-s \cdot d_{i,j}}}{\text{bias}_C \cdot \sum_{j \in C} e^{-s \cdot d_{i,j}} + \text{bias}_X \cdot \sum_{l \in X} e^{-s \cdot d_{i,l}} + K}. \quad \text{(A1)}$$

The term $d_{i,j}$ is the distance between stimulus *i* and exemplar *j*, computed as a weighted Euclidean distance:

$$d_{i,j} = \sqrt{w_x (x_i - x_j)^2 + w_y (y_i - y_j)^2}, \quad \text{(A2)}$$

where the *x*s and *y*s refer to head area and stem height measurements, respectively, and the dimension weights, w_x and w_y, sum to 1. Bias_C and bias_X are response bias parameters, and *s* is a scaling parameter that determines the sensitivity of response probabilities to a given change in distance. The probability of a category *X* or *neither* response is computed in a similar way by replacing the numerator of Equation 1 with the appropriate term from the denominator.

Without loss of generality, we factor the response threshold *K* into the product of a bias parameter and the average total similarity of each exemplar to all the others, $\overline{E}$. Thus, $K = \text{bias}_{\text{neither}} \cdot \overline{E}$, where

$$\overline{E} = \frac{1}{N} \sum_j \left(\sum_l e^{-s \cdot d_{j,l}} \right). \quad \text{(A3)}$$

N is the total number of exemplars. Higher values of $\text{bias}_{\text{neither}}$ produce more *neither* responses. There are two key advantages of this innovative treatment of the *neither* response strength. The effect of the parameter $\text{bias}_{\text{neither}}$ on goodness of fit is largely orthogonal to other parameters of the model, including the scaling parameter, *s*. As a consequence, regressions converge more quickly than with an undifferentiated *K* response strength parameter. Also, $\text{bias}_{\text{neither}}$ is independent of the number of exemplars and the average similarity among them, so meaningful comparisons can be made between best-fit values of $\text{bias}_{\text{neither}}$ obtained across different stimulus sets. In contrast, such comparisons with the undifferentiated *K* are generally not meaningful.

Also without loss of generality, we restrict the bias parameters (bias_C, bias_X, and $\text{bias}_{\text{neither}}$) to the range from 0 to 1, and require them to sum to 1 (there are only two degrees of freedom among these three parameters). Finally, we are usually interested in the relative magnitudes of bias_C and bias_X compared with each other, so we define $\text{bias}_C = \text{bias}_{C|CX}\,(1 - \text{bias}_{\text{neither}})$ and $\text{bias}_X = \text{bias}_{X|CX}\,(1 - \text{bias}_{\text{neither}})$. The parameters $\text{bias}_{C|CX}$ and $\text{bias}_{X|CX}$ determine the relative magnitudes of the two category response biases. They range from 0 to 1, and sum to 1 (having only one degree of freedom between them).

The full model defined above is fully specified with reference to four independent parameters: *s*, w_X, $\text{bias}_{C|CX}$, and $\text{bias}_{\text{neither}}$. We used this model to predict participants' classification responses, averaged across the three presentations of each flower (once in each color scheme). Thus for each of the 23 test flowers, the model predicted the proportions of category *C*, category *X*, and *neither* responses, out of three opportunities. Models were fit separately to the data for each participant by minimizing a likelihood-ratio chi-square goodness of fit statistic over the observed (*O*) and expected (*E*) response proportions:

$$G = 2 \sum O \ln\left(\frac{O}{E}\right).$$

To obtain the most accurate tests of parameter values, we eliminated parameters that did not make a significant contribution to goodness of fit. The difference in *G* obtained for a particular participant by two regression models that differ by the presence or absence of a single parameter can be tested against the chi-square distribution with one degree of freedom. Chi-square difference scores can be summed across all participants to obtain a composite statistic that can be tested against the chi-square distribution with degrees of freedom equal to the number of participants. Parameters that did not achieve a significance level of $p < .10$ on such tests were eliminated to simplify the model, as described below.

For the full model, with four free parameters for each participant, the goodness of fit was $G(164) = 604$. Setting *s* to 0 resulted in a significantly poorer fit to the data [$\chi^2(82) = 920$, $p < .001$; when $s = 0$, the parameter w_X has no effect, so there are just two degrees of freedom per participant]. This result indicates that participants' judgments were related to the distances between test and training items (i.e., those distances gave better predictions of responses than assuming that all items were treated identically). Setting w_X to 0.5 (equal weighting for both dimensions) did not significantly reduce fitness in comparison with the full model [$\chi^2(41) = 48$, $p = .22$]. Setting $\text{bias}_{C|CX}$ to 0.5 (equal bias for both categories) resulted in a significantly poorer

APPENDIX A (Continued)

fit compared with the full model [$\chi^2(41) = 81, p < .001$]. The *neither* response bias, $\text{bias}_{\text{neither}}$, is required since participants did, in fact, avail themselves of the option to respond *neither*. Taken together, these results suggest that the dimension weight parameter is not needed for our data, but that the other parameters are relevant. Accordingly, the dimension weight parameter was set to .5 in subsequent tests.

A further set of regressions tested whether the remaining parameters could be set to the median value for each participant group rather than varying freely among all participants within each group. Setting s to the median value of each group resulted in a significantly poorer fit to the data compared with including a separate free parameter for each participant [$\chi^2(39) = 58, p = .027$]. Thus, distance scaling varied within groups. Setting $\text{bias}_{C|CX}$ to the median value within each group did not result in a significantly poorer fit compared with including separate parameters for each participant [$\chi^2(39) = 41, p = .37$]. A further comparison found that setting this parameter to the overall median value ($\text{bias}_{C|CX} = 0.68$) fit the data just as well as the model that used the median of each training group [$\chi^2(1) = 0.21, p = .65$]. Setting $\text{bias}_{\text{neither}}$ to the median value of each group resulted in a significantly poorer fit to the data [$\chi^2(39) = 74, p = .001$]. This result indicates that the bias to respond *neither* varied among participants. Taken together, these results indicate that s and $\text{bias}_{\text{neither}}$ are required for each participant. $\text{Bias}_{C|CX}$ did not vary among participants.

In response to a reviewer's query, we also implemented a post hoc prototype mixture model that combined similarity to instances and to prototypes, as a weighted average, with separate scaling parameters, s, for exemplar and prototype similarities. The pure prototype model was much worse than the pure instance model (goodness of fit $G = 1{,}597$ and 693, respectively). The best-fitting mixture model fit only slightly better than the pure exemplar model, and the difference was not significant [$\chi^2(82) = 2.97, p = 1$]. Thus, similarity to prototypes had no significant effect on responses after similarity to exemplars was taken into account.

APPENDIX B
Details of Modeling for Recognition Data

In our recognition task, participants judged whether item i was old or new. The GCM predicts the probability of an *old* response on the basis of total weighted similarity of the test item to all exemplars, divided by this similarity plus a constant response threshold for *new* responses:

$$p\left(\text{response}_{\text{old}} \mid i\right) = \frac{\text{B}_{\text{old}} \sum_j e^{-s \cdot d_{i,j}}}{\text{B}_{\text{old}} \sum_j e^{-s \cdot d_{i,j}} + K}. \tag{B1}$$

The term $d_{i,j}$ is the Euclidean distance between stimulus i and exemplar j, as in Appendix A. Following Appendix A, we factor K into the product of the average similarity of each exemplar to all, $\overline{E}$, multiplied by a bias term. Thus, $K = \text{B}_{\text{new}} \cdot \overline{E}$. Higher values of B_{new} produce more *new* responses. B_{old} and B_{new} range from 0 to 1 and sum to 1 (there is one degree of freedom between them). This full model includes three independent parameters: s, w_X, and B_{new}. For each of the 31 test flowers, the model predicted the proportions of *old* and *new* responses, out of three opportunities (once in each color scheme). Models were fit separately to the data for each participant.

Again, we eliminated parameters that did not make a significant contribution to goodness of fit. Setting s to 0 resulted in significantly worse fits to the data compared with the full model [$\chi^2(64) = 159, p < .001$]. Setting w_X to 0.5 did not significantly reduce the fit of the model [$\chi^2(32) = 24, p = .85$]. B_{new} is required in order to predict any *new* judgments, so this parameter cannot be removed or set to 0. Taken together, these results suggest that the dimension weight parameter is not needed for our data, but that the other parameters are relevant. The dimension weight parameter was set to .5 in subsequent tests.

A further set of regressions tested whether the remaining parameters could be set to the median value for each participant group. The fit of the model with one s parameter for each group of participants was not significantly worse than the fit of the model with separate parameters for every participant [$\chi^2(30) = 26, p = .69$]. A further comparison found that setting s to the overall median of all participants resulted in a significantly worse fit to the data compared with separate values for each training group [$\chi^2(1) = 11.5, p = .001$]. This result indicates that the diversity manipulation affected s. The fit of the model with one B_{new} parameter for each group was significantly worse than the model with separate parameters for each participant [$\chi^2(30) = 86, p < .001$]. This result indicates that the bias to respond *new* varied among participants.

Again, we implemented a post hoc prototype mixture model that combined similarity to instances and to prototypes, as a weighted average. The pure prototype model was substantially worse than the pure instance model (goodness of fit $G = 546$ and 411, respectively). The best-fitting mixture model fit only slightly better than did the pure exemplar model, and the difference was not significant [$\chi^2(64) = 1.42, p = 1$]. Thus, similarity to prototypes had no significant effect on responses after similarity to exemplars was taken into account.

(Manuscript received December 12, 2002;
revision accepted for publication June 18, 2004.)