Journal of Experimental Psychology:
Learning, Memory, and Cognition
1996, Vol. 22, No. 2, 259-277

0278-7393/96/$3.00

Categorization and Sensitivity to Correlation

John R. Anderson and Jon M. Fincham
Carnegie Mellon University

Three categorization experiments were run in which participants saw a set of stimuli that varied on 4 continuous dimensions. Participants first categorized the stimuli and tried to predict some of the dimensions, given the values of others. Experiment 1 used iris-like stimuli based on the descriptions of R. A. Fisher's (1936) taxonomic descriptions. It showed that having participants categorize the stimuli was essential to being able to perform the prediction task and that merely observing the stimuli was not sufficient. It also indicated that participants could use within-category as well as between-category correlations for predictions. Experiments 2 and 3 used stimuli with artificial variations of values. Participants processed categories that had different within-category correlations. Participants' behavior could be predicted as a combination of sensitivity to within-category correlation and bias about the sign of the correlations. These results were fit to the rational model of categorization (J. R. Anderson, 1991) and to an exemplar model (R. M. Nosofsky, 1988).

In this article, we present research concerned with the sensitivity participants have to within-category correlations of features and try to relate the results to the rational analysis of categorization (Anderson, 1991). Most research on categorization is concerned with the correlation among features that occurs across categories. Billman (1989) argued that such across-category correlations are what drive the learning of categories. There has been some research on the learning of within-category correlation. Malt and Smith (1984) showed that participants are sensitive to the correlations that exist within natural categories. Wattenmaker (1991) showed that participants can learn such correlational structures, at least in implicit learning conditions. These studies were concerned with learning of correlations between discrete dimensions (e.g., birds that are small also sing, but birds that are large do not). There has also been a fair amount of research about how participants use the correlational structure among a set of continuous features to make predictions (multiple-cue probability learning; e.g., Hammond, McClelland, & Mumpower, 1980; Klayman, 1988). However, the role of category structure in making these predictions was not examined in the studies mentioned. Our research is concerned with learning of within-category correlations among continuous dimensions. This research originated as a test of the rational analysis of categorization and its application to a set of material described by Fisher (1936). Therefore, we next describe that theory and then the Fisher stimulus set.

John R. Anderson and Jon M. Fincham, Department of Psychology, Carnegie Mellon University.

This research was supported by National Science Foundation Grant NSF 91-08529. We would like to thank Chris Schunn for suggestions that led to Experiments 2 and 3. We also thank Stuart Elliott, Marsha Lovett, and Lael Schooler for their comments on this article.

Correspondence concerning this article should be addressed to John R. Anderson, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Electronic mail may be sent via Internet to ja@cmu.edu.

The Rational Theory of Categorization

Anderson (1990, 1991; Anderson & Matessa, 1992) proposed that human categorization could be understood as if it served the function of making optimal predictions, in a Bayesian statistical sense, about the unseen features of objects. An algorithm was developed that assumed participants would create categories and assign objects to them so as to maximize the probability of the feature structure of the objects. In particular, it is proposed that a new object with feature structure F is assigned to a category k that maximizes the probability $p(k|F)$, which is to be read as the probability that the object comes from category k given that it has feature structure F. A new category will be created if the object is not sufficiently probable given any other category. The theory treats this probability as a Bayesian posterior probability and proposes that it can be calculated according to the following formula:

$$p(k|F) = \frac{p(k)p(F|k)}{\sum_i p(i)p(F|i)}, \quad (1)$$

where $p(k)$ is the prior probability that the object comes from category k and $p(F|k)$ is the conditional probability it would display feature structure F if it did come from category k. Anderson (1990) can be consulted for a derivation of the prior probability. Our focus in this article is on the conditional probability $p(F|k)$, which turns out to be basically a measure of the similarity of the object to the category. The feature set F can be considered to be a set of separate features y_i. Given the assumption that the features within categories are independent, Anderson proposed the following formula for the calculation of $p(F|k)$:

$$p(F|k) = \prod_{y_i \in F} p_i(y_i|k), \quad (2)$$

where the $p_i(y_i|k)$ are the probabilities or densities associated

with displaying value y_i on dimension i.[1] Thus if F included the continuous feature of size y_i and the discrete feature of *sings* and if k were the category bird, then $p_i(y_i|k)$ would include the density of size y_i for birds and the probabilities that birds sing. Basically, what this equation does is calculate the probability of the feature bundle as a product of the probabilities of each feature separately. It is this independence assumption that is the focus of this article, because we consider effects of within-category correlations that violate this assumption.

We have described how categories are formed in the theory, but according to the rational theory, category formation is only a means to a prediction. Thus, the critical assumption in the theory concerns how categories formed will relate to predictions made. In particular, if one sees an object with observed feature structure F with a missing value y_i, how does one go about predicting the value of y_i? The assumption is that the following probability is calculated:

$$E(y_i) = \sum_k p(k|F)y_i(k), \tag{3}$$

where $y_i(k)$ is the value of y_i in category k. Thus, if one hears an animal singing, Equation 3 could be used to estimate its size. That is, the value on the missing dimension (i.e., whether it sings) for each category (including birds) is taken and a weighted average is formed over the categories where the weights are the probabilities that the object comes from each category. Because these weights, $p(k|F)$, are calculated according to Equations 1 and 2, the assumption of independence comes in again in the prediction process.

Note that the assumption of independence does not imply that two dimensions are independent over all objects. It only implies independence over the objects within a category. Thus, there may be an overall correlation across animals between size and ferocity. However, the assumption of independence only requires that there be no correlation within a particular species or category. Indeed, one can think of categories as capturing those regions of the object space where independence holds. They can be viewed as an attempt to convert an overall correlated relationship into a set of independent relationships. Thus, by creating multiple categories of objects one captures the potential correlations that exist in the data. There is a parallel here to latent class analysis (Lazarsfeld & Henry, 1968) that seeks to partition a set of stimuli in order to maximize correlation between sets and minimize it within sets.

The rational theory is characteristic of many theories of categorization in that it assumes within-category feature independence. Prototype theories (e.g., Reed, 1972) measure an instance in terms of its distance from a single average prototype. They do not allow for the possibility that certain patterns of features go together. However, instance-based theories (Medin & Schaffer, 1978; Nosofsky, 1988) measure a new instance in terms of its distance from other instances. Although they do not try to extract an overall relationship among features, they can nonetheless capture some correlational structure because they classify instances largely in terms of their nearest neighbors. To the extent that those nearest neighbors participate in the correlation, the predictions about a target instance will reflect this correlation.

Fisher's (1936) Iris Data Set

Fisher (1936) published a description of a set of 150 irises that came from three species. There are 50 irises from each of the species *Iris setosa, Iris versicolor,* and *Iris virginica*. Figure 1 illustrates an *I. versicolor*. Fisher published the measures of the width and length of the sepals and petals of these flowers. Table 1 displays the average values of these three species on these four dimensions. This data set has provided material for a great many efforts at categorization and clustering (e.g., Cheeseman et al., 1988). Anderson and Matessa (1992) reported the application of the categorization program based on the rational analysis to this data set. The most typical way this data set has been used is to present a categorization program with the descriptions of the 150 instances and have it try to induce a categorization of the instances. The typical result of such categorization efforts (including our own program) is that the program forms a category that corresponds to the *I. setosa* but has great difficulty in separating out the *I. versicolor* from the *I. virginica*. A frequent behavior is that a category is formed that corresponds to the *I. versicolor* and the smaller *I. virginica* and that a separate category is formed for the larger *I. virginica*. Another frequent outcome is that the *I. versicolor* and *I. virginica* are merged into a single category.

Anderson and Matessa (1992) created caricatures of these flowers, such as the example in Figure 2, and asked participants to sort them into categories. These stimuli only varied in the length and width of their artificial sepals and petals. Participants displayed the same behavior as the program, separating out the *I. setosa* but showing similar confusions between *I. versicolor* and *I. virginica*. Interestingly, botanists (Mathew, 1981) are not totally in agreement as to whether *I. versicolor* and *I. virginica* should be treated as separate species. As we will see, if participants are told what categories the flowers come from, they can learn to reproduce the conventional botanist categorization. It is also the case that Anderson and Matessa's program, given the category's labels, will also reproduce the botanist categories. This is because the labels are another dimension to predict, and greater predictability is now achieved if the internal categories correspond to the botanist's categories. So, in essence, the labeling can determine the categories with which a participant or the program will perceive the world.

The principal purpose of the rational theory is not to extract botanists' current theory of the species structure of irises. Rather, it is to be able to predict unseen dimensions given seen dimensions. For instance, if the program is given petal length and width as well as sepal width, how well can it do at predicting sepal length? It turns out that the ability of our rational program to do this prediction was only affected a little by whether it used its own categories or those of the botanists (Anderson & Matessa, 1992). Also, its predictions were much better than chance. The average squared error of prediction of the program was 0.26 cm^2, whereas if the average value was chosen it would have been 1.20 cm^2. Thus, the category structure that the program induced was capturing a lot of the

[1] Anderson and Matessa (1992) can be consulted for a Bayesian analysis of how these conditional probabilities should be estimated.

predictable variance. The program did somewhat better if it used the botanist categorization (0.20 cm^2 error), which can be seen as some validation of the conventional categories.

However, it turns out that a more accurate approach to categorization is simply to calculate linear regression equations over all categories, predicting the values of a fourth dimension given the other three. The mean error of this prediction scheme was 0.09 cm^2. The reason for this is that there were strong correlations among the dimensions, not only between categories but within categories. Table 2 displays the overall correlation matrices as well as the within-category correlation matrices. As can be seen overall, there were fairly high intercorrelations among petal width, petal length, and sepal length but correlations with sepal width were weak. The within-category correlations were similar but not identical. For instance, although sepal width did not appear to enter into any positive correlations overall, it generally had positive correlations within categories. The existence of these within-category correlations is a direct violation of the assumptions of the original rational model. To correct for this problem, Anderson and Matessa (1992) created a rational algorithm that calculated Bayesian estimates of intracategory correlations and used these for prediction. This algorithm slightly outperformed the regression algorithm, with a mean error of 0.08 cm^2.

Figure 1. An example of *Iris versicolor* from Dana (1893).

Table 1
Average Values (in Centimeters) and Standard Errors of the Three Species of Irises

Dimension	*Iris setosa*	*Iris versicolor*	*Iris virginica*
Petal width	0.25 ± 0.11	1.33 ± 2.00	2.03 ± 0.28
Petal length	1.46 ± 0.17	4.26 ± 0.47	5.55 ± 0.55
Sepal width	3.43 ± 0.38	2.77 ± 0.31	2.97 ± 0.32
Sepal length	5.01 ± 0.35	5.94 ± 0.52	6.59 ± 0.64

This raises the question of how well participants do at learning the correlational structure of such stimulus sets. As noted in the introduction, past research has indicated that at least in some circumstances, participants can pick up on correlations between discrete features. It turns out that our rational algorithm does this also. For instance, the rational model (Anderson, 1991) was able to simulate the research of Medin, Altom, Edelson, and Freko (1982), which showed that participants could pick up a correlation between a pair of binary features (if there was a 0 on one dimension there would be a 0 on the other; if there was a 1 on one dimension there would be a 1 on the other). Our rational categorization program (Anderson, 1991) also picked up on this correlation because it created separate categories to record the two possible values of the correlation (i.e., a 0–0 category and a 1–1 category). The general strategy of a separate category for each feature pair is not practical in the case of continuously varying dimensions because one would need a separate category for each stimulus (reducing the rational model to the exemplar model—see Nosofsky, 1991). Thus, the continuous stimuli in the iris set offer an opportunity to test whether participants are sensitive to within-category correlations in a way that the discrete stimuli of past experiments did not. The basic purpose of the first experiment was to see if humans are sensitive to the within-category correlations in Fisher's (1936) material.

Experiment 1

We wanted to contrast various conditions of exposure to the original Fisher (1936) dimensions presented in our format (see Figure 2). In one condition, participants were trained to categorize the stimuli according to the official botanist categorization. However, our earlier pilot work showed that participants did not naturally identify the botanist categories but rather created categories more like those identified by the machine learning programs. We were interested in whether there would be differential sensitivity to correlations when participants were trained on categories or were free to make

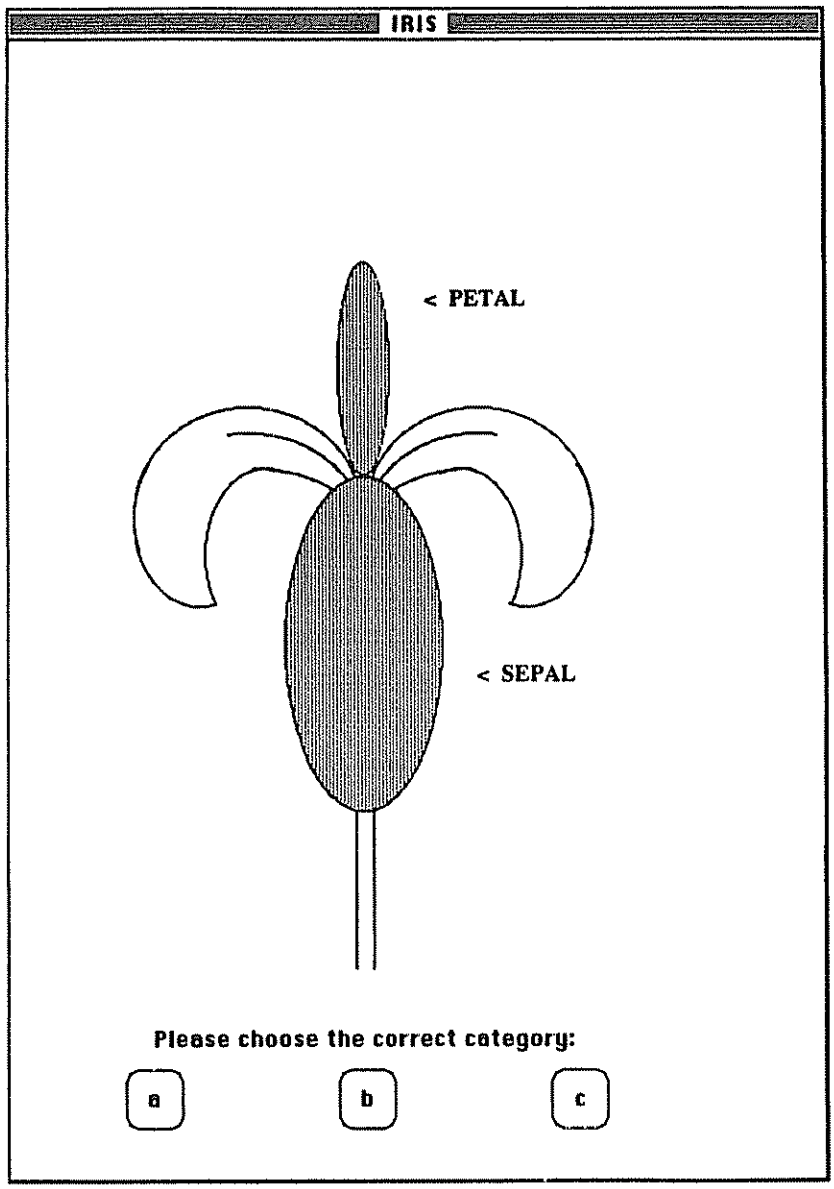

Figure 2 Schematic iris used in Experiment 1.

up their own categories. Therefore, we created a second condition in which participants created their own categories. Wattenmaker (1991) found that participants were sensitive to correlations only when they were not explicitly categorizing the stimuli. Therefore, we created a third condition in which participants were exposed to the stimuli but did not have to learn to make category assignments and instead had to rate how much they liked the stimuli. Finally, to have a reference condition, we created a fourth condition in which participants had no prior exposure to the stimuli. In all conditions, participants transferred to a test phase where they saw our irises with one or two dimensions missing and had to reproduce the missing dimension(s). Because sepal width did not vary much, we did not require participants to reproduce it, but we tested all combinations that had one or two of the remaining features missing.

Method

Participants. There were 10 participants in each of the following four groups: botanist categorization, self-categorization, likeability ratings, and no prior exposure. Participants were recruited from the Carnegie Mellon University undergraduate population and were paid $5.00 for their involvement (which lasted less than an hour).

Materials. The materials were the caricature irises based on the Fisher (1936) dimensions (see Figure 2). The sepals and petals were ovals drawn to have the width and length of the original Fisher stimuli. To fulfill the structure of the experimental design, we only needed 48 stimuli from each category. Therefore, 48 members were randomly

Table 2
Intercorrelation Matrixes for Fisher's (1936) Stimuli

Dimension	Dimension			
	PX	PY	SX	SY
Overall				
PX	—			
PY	.96	—		
SX	−.37	−.43	—	
SY	.82	.87	−.12	—
Within *Iris setosa*				
PX	—			
PY	.33	—		
SX	.23	.18	—	
SY	.28	.27	.74	—
Within *Iris versicolor*				
PX	—			
PY	.79	—		
SX	.66	.56	—	
SY	.55	.75	.53	—
Within *Iris virginica*				
PX	—			
PY	.32	—		
SX	.54	.40	—	
SY	.28	.86	.46	—

Note. PX = petal width; PY = petal length; SX = sepal width; SY = sepal length.

chosen from the 50 for each species. A random 24 from each category were presented during study, but all 48 were used during test. At study, the stimuli were presented with all dimensions present. At test, the stimuli were presented with one or two of the dimensions set to zero length. Figure 3 shows a test stimulus with the sepal length and petal width zeroed out. Pressing the + key associated with the sepal length would increase it in the vertical dimension and the horizontal would stay fixed; pressing the + key associated with the petal width would fatten it but keep it at a fixed length. There were six possible ways of zeroing out one or two of the dimensions of sepal length, petal length, and petal width. To instantiate each of these six test types, 4 studied stimuli and 4 nonstudied stimuli were randomly chosen from each species. Thus, the test consisted of 144 items: 6 (ways of testing) × 3 (species) × 4 (instances) × 2 (studied vs. nonstudied). All materials were presented by means of a Macintosh II computer.

Procedure. In the botanist categorization condition (which is illustrated in Figure 2), participants were given three categories (a, b, and c) and were required to assign objects to these categories in a way such that one of the letters would correspond to *I. setosa*, one to *I. virginica*, and one to *I. versicolor*. Participants' initial placement of stimuli determined which of a, b, or c corresponded to the categories. After that, they were given feedback after every error as to the correct categorization of the flower. Participants made as many passes through the material as were necessary for them to correctly categorize at least 80% of the stimuli. In the self-categorization condition, participants were allowed to use up to five categories denoted by the letters *a* through *e*. They made however many passes through the stimuli as were required so that their assignments in the final pass overlapped at least 80% with their assignments in the previous pass. In the likeability ratings condition, participants had to decide whether they found the particular flower pleasing, displeasing, or neutral. They made however many passes through the stimuli as were required so that the evaluations in the final pass overlapped 80% with the evaluations in the previous pass. In the no prior exposure condition, there was no study phase. In the other three conditions, participants were not warned of the upcoming test phase and presumably had no reason to suspect that such a test phase would occur.

In the test phase, participants were instructed to size the stimuli so that they were like the stimuli they had studied. The stimuli were presented and the participants pressed a + key associated with that dimension to increase it and a − key to decrease it. The keys associated with fixed dimensions were shaded gray and could not be used. When participants were satisfied with their reproduction, they pressed an "OK" button. The order of the stimuli in all phases was randomly determined. All of the participants' choices were recorded for later analysis.

Results

The mean number of passes through the study stimuli to reach criterion was 3.9 in the botanist categorization condition, 2.3 in the self-categorization condition, and 2.8 in the likeability rating condition. An overall test for significance among the conditions was only marginally significant, $F(2, 27) = 2.89, p < .1, MSE = 2.32$, but a specific contrast that asked if the botanist categorization condition was longer than the average of the other two was significant, $t(27) = 2.08, p < .05$.

For the data from the test phase, the absolute deviations of the participants' reproductions from the true values were measured. These data were subjected to an analysis of variance (ANOVA) in which the variables were study condition (four levels), dimension to be predicted (three levels), dimensions available for predicting those dimensions (three levels), species of iris (three levels), and whether the iris had been studied or not. With respect to dimensions available for predicting, we classified these as either three dimensions or two dimensions. In the two-dimension case, we classified these according to whether the highest correlate of the to-be-predicted dimension was present (on the basis of Table 2). When predicting petal length, the higher correlate was petal width and the lower correlate was sepal length. When predicting petal width, the higher correlate was petal length and the lower correlate was sepal length. When predicting sepal length, the higher correlate was petal length and the lower correlate was petal width. So, the three levels for the available dimensions variable were three dimensions, two dimensions including the higher correlate, and two dimensions including the lower correlate.

All main effects were significant, except whether the items were old or new. Some of the main effects simply reflected differences in materials. Participants were more accurate in predicting petal width (mean error 0.73 cm) than petal length (mean error 1.53 cm) or sepal length (mean error 1.56 cm)—$F(2, 72) = 29.96, p < .0001, MSE = 2.69$. Participants were most accurate with *I. setosa* (mean error 1.07 cm), next most accurate with *I. versicolor* (mean error 1.22 cm), and least accurate with *I. virginica* (mean error 1.53 cm)—$F(2, 54) = 21.67, p < .0001, MSE = 0.94$. Both the dimension and category effects were such that participants made larger errors in reproducing larger objects.

Figure 4 shows the effect of prior exposure. Participants were equally accurate whether they were trained with the experimenter's categories or used their own, but they were much less accurate when they had given likeability ratings or had no prior exposure, $F(3, 36) = 8.67, p < .001, MSE = 8.58$.

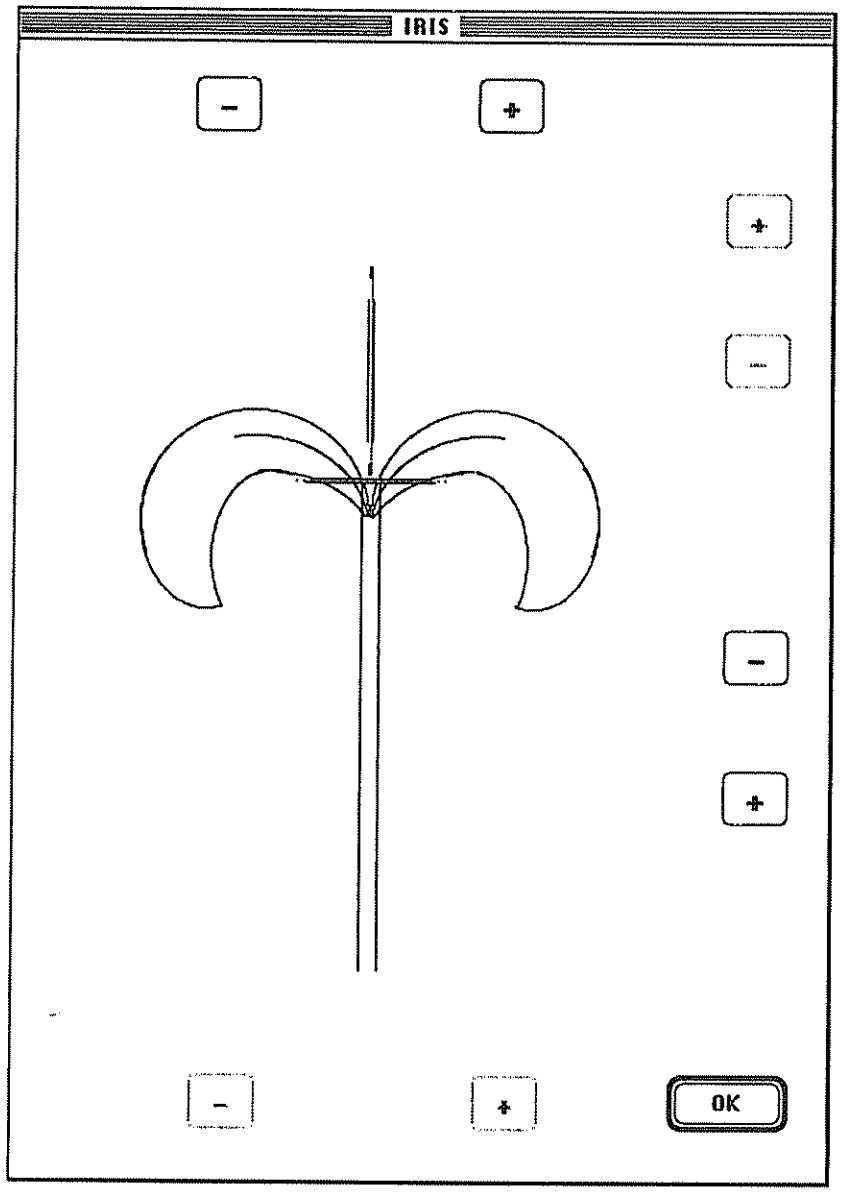

Figure 3. Example test stimulus. The participant had to reproduce width of the petal and length of the sepal.

The difference between likeability ratings and no prior exposure was not significant, $t(36) = 0.55$. Thus, participants were able to do much better than chance (no prior exposure) at predicting the dimensions given a categorization task but not given an exposure task. This shows that even though participants were not explicitly predicting widths and lengths and were not expecting to be so tested, there was something about forming a category structure that enabled them to make such predictions. This is certainly in keeping with the rational analysis of categorization behavior.

Overall, there was a significant effect of available dimensions, $F(2, 72) = 15.36$, $p < .0001$, $MSE = 0.31$, and an interaction with study condition, $F(6, 72) = 3.24$, $p < .01$, $MSE = 0.31$, such that the effect of the dimensions showed only in the self-categorization and botanist categorization conditions. Figure 5 displays the effect of dimensions available averaged over the botanist categorization condition and the self-categorization condition. Participants were more accurate in the presence of the high correlate than in the presence of just the low correlate. Further, adding the low correlate to get the three-dimension condition did little to enhance prediction.

Rational model for the botanist category condition. The effect of available dimensions indicates that participants were sensitive to the correlational structure of the stimuli. However, it is unclear how much sensitivity there was to within-category correlation. It could just be that having available the high correlate increased the accuracy of categorization and that

participants were simply using category averages to predict the target dimensions. To determine within-category effects, we have to know where the category boundaries are. We only know this precisely and consistently for the participants in the botanist categorization condition. Therefore, subsequent analyses focus on that condition.

We broke the flowers in each species into large and small sizes on the basis of the sum of the four dimensions. This gave us six subcategories that ranged from small *I. setosa* to large *I. virginica*. Figure 6 compares the mean participant reproductions with the true values and with the predictions of two models to be described. In this figure, we are averaging over the three conditions of available correlates but we do present the values for each of the three predicted dimensions. As can be seen, participants were sensitive to the within-category size variation as well as to the between-category size variation, although participants tended to underproduce the true variation everywhere. Averaged over the three dimensions, the true values varied from 2.13 cm for small *I. setosa* to 5.04 cm for large *I. virginica*, but the reproduced values varied from 2.40 cm to 4.53 cm. The actual correlation between the participant averages and the true values in Figure 6 was .987. So, even though participants had no expectation that they would be tested on their ability to reproduce the values in the irises, the categorization training led to quite high accuracy in mean reproduction.

We tried to fit the rational model to the data in the botanist categorization condition based on Equations 1–3. The data fit were the 54 averages defined by crossing the 18 conditions in Figure 6 with the three possibilities for available correlates.

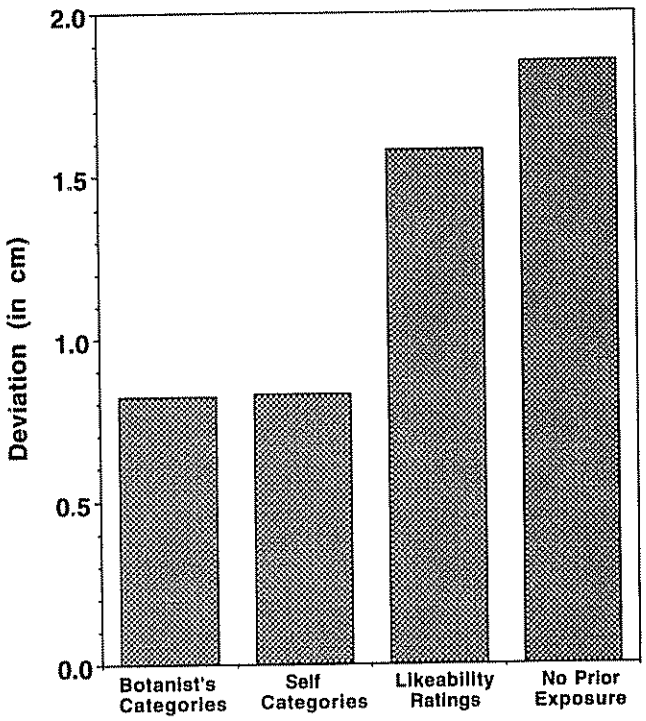

Figure 4. Mean centimeters of error in reproduction as a function of the training condition.

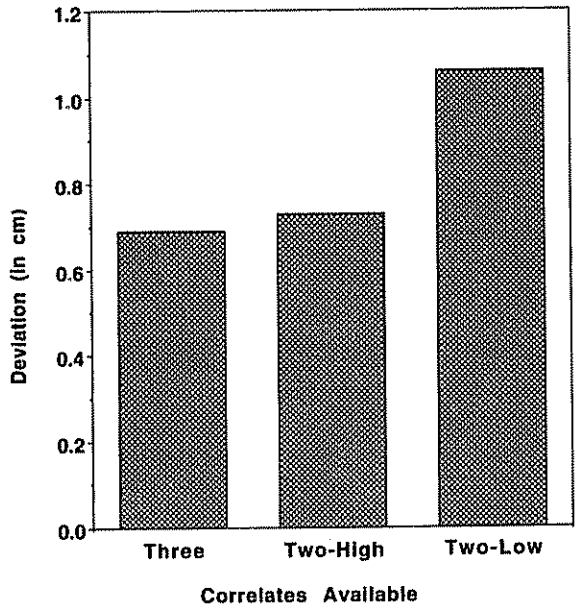

Figure 5. Mean centimeters of error as a function of the correlates available. These means are just from participants in the training conditions of botanists categories and self-categories.

Equation 2 was used to assign an overall measure of similarity of each average stimulus to each category. Thus, for each condition of available correlates, we calculated the product of the similarities to the category means of the values y_i on the two or three dimensions available. According to the rational model, the $p_i(y_i|k)$ are probability densities calculated according to a Bayesian analysis assuming a normal prior distribution for the mean and an inverse chi-square distribution for the variance of the y_i (see Anderson, 1991; Anderson & Matessa, 1992). In the model we fit, we estimated the mean of the prior distribution of the mean and the mean of the prior distribution of the variance. The Bayesian analysis involves developing a posterior distribution of the mean and the variance as a weighted average of the prior and the empirically observed mean and variance for that dimension and that category. The empirical means and variances are weighted by their numbers (24) of observations, but the Bayesian analysis requires estimating a weighting of the priors. This parameter is essentially a measure of the number of observations a prior is worth. Thus, we estimated three parameters: a single prior mean, μ, for all dimensions and categories; a single prior variance, σ^2, for all dimensions and categories; and a single weighting, λ, of this prior mean and variance. If μ_k is the weighted mean for category k and if σ_k^2 is the weighted variance for the category, the $p(y_i|k)$ are calculated according to a normal distribution:[2]

$$p(y_i|k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(y_i - \mu_k)^2/\sigma_k^2} \quad (4)$$

[2] As discussed by Anderson and Matessa (1992), the more accurate Bayesian model assumes a t distribution, but the difference with the normal is very minor.

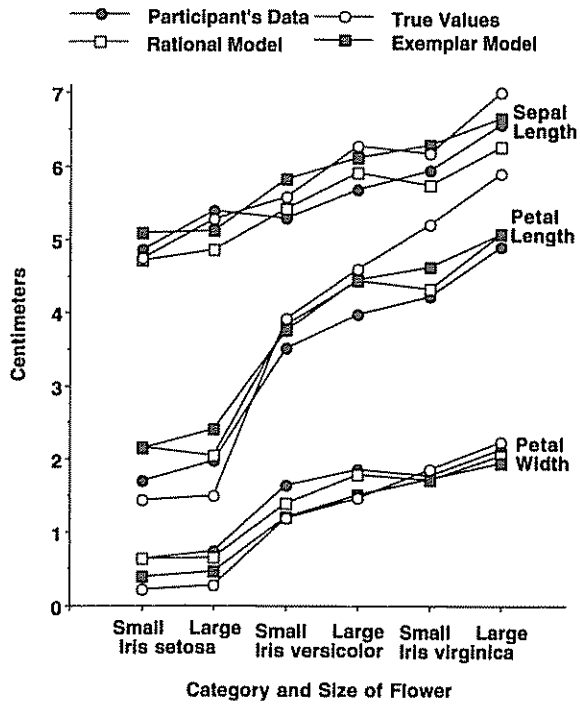

Figure 6. Absolute size of dimensions as a function of the size of the category from which the flower was taken. Separate functions are plotted for the three dimensions to be predicted. For each dimension the figure compares the true values, the participant values (in the condition of botanist categories), and the predictions of the rational model and the exemplar model.

To be totally precise, we would have had to make predictions for the 40 (10 participants × 4 observations) stimuli in each condition and then average them. However, we simply made predictions for average stimuli that were defined by the average values over the 40 stimuli. These predictions were made by weighting the category means according to Equation 3, using weights calculated by Equations 2 and 4 and using category means that were the posterior means of the Bayesian estimates.[3]

The best fitting values of these parameters were $\mu = 2.4$ cm, $\sigma^2 = 4.99$ cm^2, and $\lambda = 2.75$ observations. With these parameters we achieved a total squared deviation between the 54 predictions and observations of 6.80 cm^2, which is equivalent to an R^2 of .966. However, with a mean square error of the means of 0.051 cm^2, this is a significant difference, $F(51, 477) = 2.61$, $p < .001$. One dimension of the problem is that this model is not capturing enough of participants' sensitivity to within-category variance. The difference between participants' reproductions for larger and smaller members of a category was 0.40 cm (the actual difference in the flowers was 0.45 cm). The model predicts some sensitivity because, when it is predicting for the larger half of a category, it will tend to give larger categories higher weightings in Equation 3. However, it only predicts a 0.13-cm difference. Also, recent research by Murphy and Ross (1994) has suggested that participants do not weight multiple categories in making a prediction but simply select the dominant category. This casts doubt on the validity of using category weightings as a basis for predicting sensitivity to within-category correlation.

As mentioned earlier, Anderson and Matessa (1992) developed a version of the rational model that used within-category correlations. We tried a reduced version of this model that just used the largest correlation to adjust its prediction. So the mean for dimension i for category k was

$$f_i(y_i|k) = m_i + \alpha \cdot b_{ji} \cdot d_j, \tag{5}$$

where m_i was the posterior mean for the to-be-predicted dimension i (calculated as above), d_j was the deviation of the subcategory mean from the category mean for the predicting dimension j (negative for small, positive for large), b_{ji} was the estimated slope of the predicted dimension i against the predicting dimension j derived from the within-category correlation,[4] and α was an attenuation factor. The attenuation factor can be interpreted as reflecting the outcome of weighting a prior of no correlation with the observed correlation. This model is a four-parameter model because it has one additional parameter to reflect the attenuation factor. (If the attenuation factor were zero, this would be the same as in the previous model.) The best fitting values of these parameters were $\mu = 2.45$ cm, $\sigma^2 = 4.38$ cm^2, $\lambda = 2.94$ observations, and $\alpha = 0.67$. These are all reasonable values with the mean and variance within the range of the stimuli, the prior weight within the magnitude of the observations, and the attenuation factor positive and less than one. With these parameters, we achieved a total squared deviation between predicted and observed values of 5.83 cm^2, which is equivalent to an R^2 of .971. With the mean square error of the means of 0.051 cm^2, this is still a significant difference, $F(50, 477) = 2.29$, $p < .005$. The reduction in deviation is also quite significant, $F(1, 477) = 19.02$, $p < .001$. This model predicts a 0.35-cm difference between the reproductions of the small and large members of a category, which is close to the participants' difference of 0.40 cm. It is this four-parameter model that is plotted in Figure 6.

Exemplar model for the botanist category condition. Elliott and Anderson (1995) have developed an application of the exemplar model of Nosofsky (1986; Nosofsky, Clark, and Shin, 1989) for predicting the data in such an experiment. This model predicts that participants use the actual exemplars to make predictions. The predicted value is

$$E(y_i) = \sum_j s_j x_{ij} \Big/ \sum_j s_j, \tag{6}$$

where the summations are over the individual study instances j, s_j is the similarity of the study instance to the test stimulus, and

[3] The rational model allows for a "coupling" parameter that determines whether a stimulus gets placed into a new category. For simplicity, we ignored this parameter and so only considered the possibility that the instances came from one of the existing categories. The effect of this approximation was minor because there were many instances seen per category, which would make the probability of a new category low.

[4] The b_{ji} are simply the empirical slopes manifested in the Fisher (1936) stimuli and do not involve estimating any free parameters.

x_{ij} is the value of that study instance on dimension i. This is similar to Equation 3 for the rational model except that individual instances (rather than category means) are being weighted by their similarities to the test item. Similarity is calculated as a product of similarities on each dimension (similar to Equation 2 for the rational model):

$$s_j = \prod_i e^{-w_i|y_i - x_{ij}|}, \quad (7)$$

where the product is over the observed dimensions, w_i is the attentional weight given to dimension i, y_i is the observed value on dimension i, and x_{ij} is the value of study instance j on dimension i. This model is similar in many ways to the rational model except that it uses individual study instances rather than categories. It uses an exponential function to weight the differences rather than the sigmoidal function in the rational analysis (Equation 4), but this is a minor difference for current purposes. The significant differences come from use of specific instances. The lack of any categories implies that it is not possible to calculate within-category correlations as in the extended rational model. It is an open question whether correlations are necessary. In particular, as noted in the introduction, by recording each example the model allows itself to empirically represent the correlation.

We fit the model using all 150 Fisher (1936) irises as past instances but, as in the rational model, predicted the 54 aggregate test stimuli. As is typically done, we allowed ourselves to fit the model estimating separate attentional parameters w_i for each of the dimensions. The values estimated were 1.48 for petal width, 3.60 for petal length, 0 for sepal width, and 1.32 for sepal length. With these parameters we achieved a total squared deviation between prediction and observed of 7.96 cm^2, which is equivalent to an R^2 of .960. With the mean square error of the means of 0.051 cm^2, this is still a significant difference, $F(50, 477) = 3.12$, $p < .001$. It is the Nosofsky (1986) four-parameter exemplar model that is plotted in Figure 6. The model's predictions are slightly worse than the rational model. Its major deficit is that it is not able to reproduce the average values generated by participants for the three dimensions. It underpredicts the values of the participants' reproductions for the petal width by an average of 0.26 cm, and it overpredicts the petal and sepal length reproductions by an average of 0.30 cm. It does not have the advantage of the rational model of weighting the empirical mean by a prior value. It would be easy to extend the exemplar model to have such priors.[5] We did not do so because our main concern was whether the exemplar model could produce the sensitivity to within-category correlation. On this score, the model has little difficulty with the within-category correlation. It predicts participants will produce a 0.30-cm difference between the small and large members of the category, whereas they produced a 0.40-cm difference.

Conclusions

The rational model needed to be extended to estimate within-category correlations to be able to capture participants' sensitivity to within-category correlations. This is something that the exemplar model does naturally. Later experiments provide data for discriminating between the extended rational model and the exemplar model.

The results involving the training conditions are informative. The fact that the likeability condition was no different than the no prior exposure condition and that both were much worse than either of the categorization conditions seems to indicate that mere exposure is not enough to get participants to encode the dimensional relationships and that there is something special about categorization. The fact that there was no difference between the botanist categorization condition and the self-categorization condition might seem problematical for the rational model because participants tended to collapse the *I. versicolor* and *I. virginica* together. However, if they were storing within-category correlations, their ability to predict stimulus dimensions would not have been nearly so dependent on their coming up with a particular categorization. Indeed, by adding within-category correlation monitoring, we make the model much less dependent on coming up with a good categorization.

Curiously, Wattenmaker (1991) found different results. His participants were able to detect the correlation between discrete values (e.g., has a dog and drinks carbonated soda) much better when they just studied the stimuli and did not categorize them. Wattenmaker argued that there were too many dimensions to monitor when participants were consciously forming categories. Perhaps the difference in our case turns on the relative naturalness of the correlations involved, or perhaps it has to do with the fact that we were using continuous dimensions. In any case, Wattenmaker's research establishes that it is not always necessary to categorize in order to pick up on within-category correlations.

Experiment 2

Experiment 1 provided evidence that participants are sensitive to within-category correlations in naturally occurring categories. We noted that the original rational model (without correlation monitoring) showed some sensitivity. This is because it produced different weightings of the three categories (see Equation 3) when predicting small versus large members of any category. Although the rational model did not produce enough sensitivity, the possibility remains that some weighting scheme could. We also noted in the introduction that a multiple-regression scheme that did not use categories at all did a good job of predicting the dimensions. This second experiment was designed to produce better evidence that participants were using both a category structure and within-category correlation. This experiment offers the possibilities of ruling out the weighted-category approach or the multiple-regression approach. As will be shown, it also yields data that discriminate against the exemplar model.

We abandoned the pseudonatural stimuli of the first experiment and went to using truly artificial materials that were

[5] Indeed, Elliott and Anderson (1995) describe a version of the exemplar model that involves the use of prior means. We also did not use this extension just to keep the number of parameters the same between the rational and exemplar model.

based loosely on the Fisher (1936) stimuli. One potential criticism of Experiment 1 was that participants may not have treated width and length as independent dimensions. Indeed, some participants reported processing the stimuli in terms of shape and size. To deal with this, we decided to transform the dimensions into size and shading. All petals had the same shape and varied in size (area) and shading.

Figure 7 illustrates the design of the experiment, which involved training participants to classify the stimuli into two categories. Two of the dimensions were designated as defining dimensions, and their use is illustrated in Figure 7a. Two categories were defined by conjunctions of values on these dimensions. All instances on one end of the scale for the two dimensions were in one category, and all instances on the other end of the scale were in another category. Note that, overall, there was a strong correlation between the two dimensions but that within each category there was no correlation between the dimensions. Figure 7b illustrates the structure of the two nondefining dimensions. Overall, there was no correlation between the dimensions, but within one category there was a positive correlation, and within the other category there was a negative correlation. Thus, if participants were able to predict anything about these two dimensions it would be because they were sensitive to the within-category correlation.

Method

Participants. Forty-nine Carnegie Mellon University undergraduates were paid $6.00 each to participate in an experiment that lasted less than 2 hr. In addition, participants received bonus pay of up to $6.00, depending on performance. There were 24 participants in the self-categorization condition and 25 in the trained-categorization condition.

Materials. The stimuli were constructed as in the previous experiment. However, the combinations of values for each dimension were completely artificial. The distributions of values over the dimensions were perfectly uniform. The range of shading for both sepals and petals was from 3 to 127 in the Macintosh gray scale (near white to black). The range of size values was 2.6 to 7.2 cm long, and width was one half of the length. We generated 144 values each for shading of petal, shading of sepal, size of petal, and size of sepal uniformly on these dimensions. The defining dimensions could be either the size dimensions or the shading dimensions. When size was the defining dimension, we combined the sepal and petal size to create either a positive correlation or a negative correlation. In the case of a positive correlation, the two categories would be (a) large sepals and large petals and (b) small sepals and small petals. In the case of a negative correlation, there would be (a) large sepals and small petals and (b) small sepals and large petals. Similarly, when shading was the defining dimension, the correlations could be positive or negative. Thus, there were four possible conditions of defining dimensions. For the nondefining dimensions, there was one perfect positive correlation and one perfect negative correlation. There were two ways to assign these to the two categories. Crossing these two ways with the four ways of creating defining dimensions yielded eight stimulus conditions.

Procedure. In the trained categorization condition, the participant was trained to categorize the stimuli into one of two categories according to the same procedure as that used in the previous experiment. In the self-categorization condition, participants were free to develop their own category structure using from one to five categories. Participants were trained to categorize one set of 36 stimuli and then another set of 36 stimuli. Each set of 36 contained 18 stimuli randomly selected from each category. After selecting the category for a stimulus, the participant was presented with a new window that contained a prototype iris—that is, an iris that had the overall median value on all dimensions. Participants were required to adjust the value on each dimension in order to reproduce the flower they had categorized immediately before. The participant was instructed to

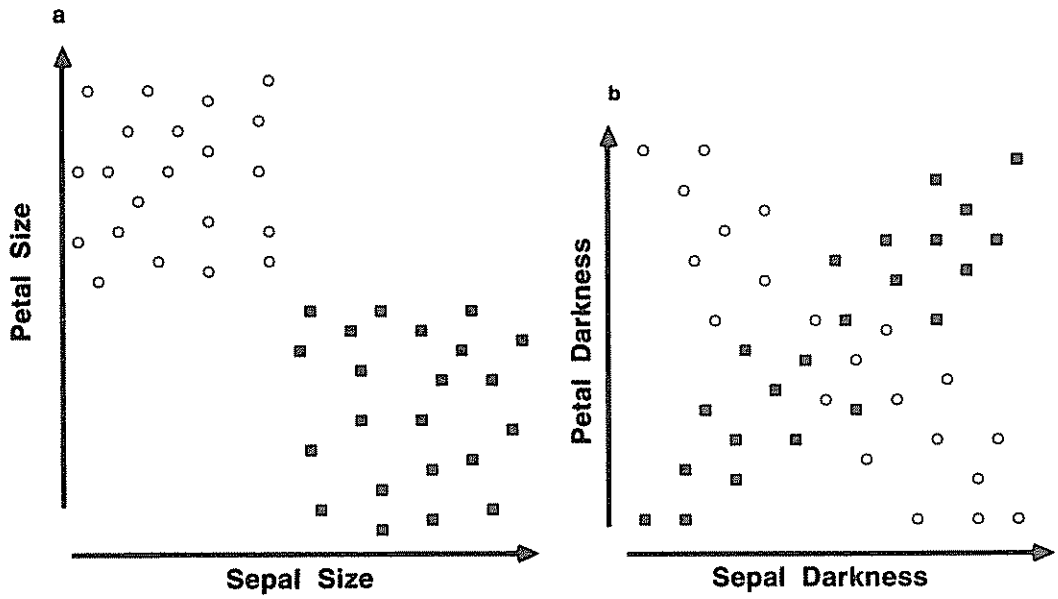

Figure 7. Example stimulus structure in Experiment 2: (a) two defining dimensions and (b) two nondefining dimensions. Open circles represent one category, and filled squares represent another condition.

perform this task well in order to maximize accuracy. In this experiment, a point score was displayed at the top of the screen. The greater the accuracy of the reproduction, the more the point score was increased. Participants were told that the final point score would be used to determine the number of bonus dollars they would receive upon completion of the experiment. This reproduction procedure was added to get participants to pay attention to all dimensions (which they had not been doing in a pilot experiment).[6] After reaching a criterion of 80% correct classification on the 36 stimuli, participants had to achieve a similar level of accuracy on another 36 stimuli.

After completing the training phase of the experiment, participants were transferred to the prediction phase. They were presented with three of the dimensions and had to predict the fourth. They were tested with all 144 stimuli divided into eight conditions defined by category and dimension to be predicted.

Results

To make shading and size comparable, we recoded them on a 0–1 scale. If M was the maximum value on the scale and m was the minimum, and x was the value reproduced by the participant, we coded this as $(x - m)/(M - m)$. We classified the participant's response according to the value of the correlated or cue dimension. If the participant was reproducing a defining dimension, the cue dimension was the other defining dimension. Similarly, if the participant was reproducing a nondefining dimension, the cue dimension was the other nondefining dimension. This cue dimension was divided into quintiles and we aggregated the data from each quintile; that is, we averaged the values reproduced for the stimuli with the smallest fifth of cue dimension, the next smallest fifth, and so on.

We performed an ANOVA on the recoded values in which the variables were training (experimenter categories vs. self categories), dimension that defined the categories (size or shading), direction of correlation in the case of the defining dimension (positive or negative), predicted dimension (defining, nondefining with positive correlation, nondefining with negative correlation), and quintile of the cue dimension. There were no significant effects of training or defining dimension (size or shading) allowing us to collapse over these two variables. There was a highly significant interaction of the three remaining variables (Direction of Defining Correlation × Predicted Dimension × Quintile), $F(8, 376) = 18.08$, $p < .0001$, $MSE = 0.023$, as well as numerous lower-order effects: predicted dimension, $F(2, 94) = 3.95$, $p < .05$, $MSE = 0.017$; quintile, $F(4, 188) = 3.92$, $p < .01$; $MSE = 0.032$; Quintile × Defining Correlation, $F(4, 188) = 26.25$, $p < .0001$, $MSE = 0.032$; and Predicted × Quintile, $F(8, 376) = 13.17$, $p < .0001$, $MSE = 0.023$.

Figure 8 provides the relevant display of these effects. Here we have broken up the dimensions being predicted into four conditions. For the nondefining dimensions, there were the positive and negative correlations again. However, we have also divided up the predictions on the defining conditions. Those participants for whom the defining dimensions were positively correlated are plotted separately from those participants for whom the dimensions were negatively correlated. We have plotted the standardized values as a function of quintile. As can be seen, we have positive functions in the presence of positive correlations and negative functions in the presence of negative correlations. For the nondefining dimensions, the positive function was much steeper than the negative function.

The dimension of experimenter-defined versus participant-defined categories did not interact with anything. However, none of the 24 participants who did self-categorization used two categories, let alone the experimenter's categories. Two participants used one category, 10 used three categories, 4 used four categories, and 8 used five categories. It was not always possible to see a basis for the participant's categorization. Ten participants had one category that corresponded to the experimenter's category in which there was a positive within-category correlation. These 10 participants broke the experimenter's category with a negative correlation into two—one reflecting the positive–negative values on the nondefining dimensions and the other the negative–positive values. We call these the three-way participants. Five participants had four categories that broke both of the experimenter's categories in half according to the extremity of the values on nondefining dimensions. We call these the four-way participants. Then there were the 2 one-way participants with just one category. Finally, there were 7 participants whose classification defied description in terms of the experimenter's categories. The three-way and four-way participants sometimes had additional categories that also defied description. We did analyses of whether there were any interactions between these four types of participants and performance but failed to get any significant interactions. So, once again we had a failure of the nature of categorization training to have an impact on the prediction results.

Analysis of experimenter's category condition. Although the data did not seem to vary as a function of condition, we decided to focus on the data in the experimenter's category condition because in this condition all 25 participants used the same categorization. Again we tried to fit the rational model to the data (i.e., the 20 data points in Figure 8, but only for participants using the experimenter's categories). We fit the same model to the data as in Experiment 1, except that instead of using the attenuation factor to convert the empirical correlations into the posterior correlations, we used the following Bayesian formula from Anderson and Matessa (1992, derived from Box & Tiao, 1973) for mixing observed correlation, r, with prior correlation, π, to estimate a final correlation, $\hat{r}$:

$$\hat{r} = \tanh\left[\frac{\lambda_\pi \tanh^{-1}\pi + 18\tanh^{-1} r}{\lambda_\pi + 18}\right], \qquad (8)$$

where λ_π is the weighting of the prior correlation. The empirical correlation was .9 for the nondefining dimensions and 0 for the defining dimensions.[7] The parameters were

[6] The fact that participants did not naturally encode all dimensions could be accommodated by adding a selective attention process as in Kruschke (1992) or Nosofsky (1986, 1988).

[7] The actual correlation was perfect, but we used .9 because $\tanh^{-1}$ is not defined for 1. This amounts to asserting that there was some noise in the perception of the correlation.

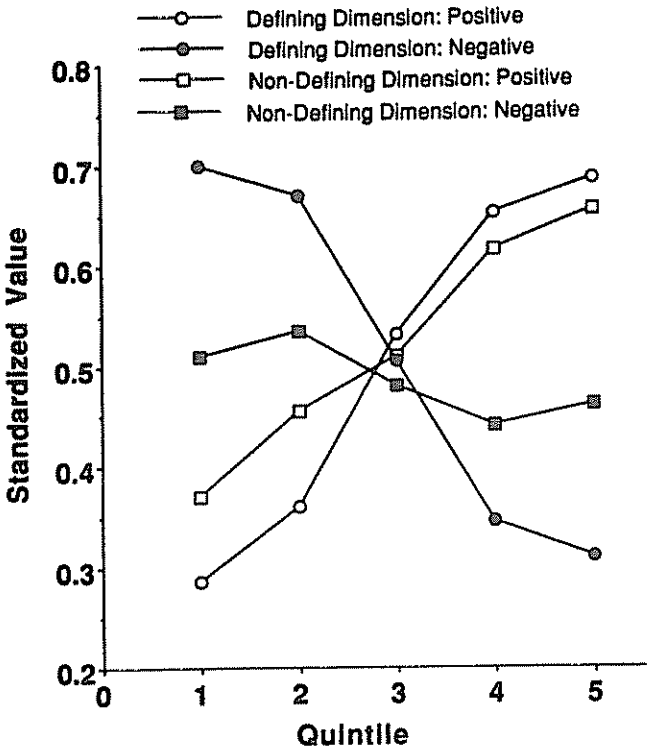

Figure 8 Standardized score as a function of the value on the cue dimension and type of dimension being predicted.

estimated as $\mu = 0.509$, $\sigma^2 = 0.371$, $\lambda = 3.26$ observations, $\pi = .158$, and $\lambda_\pi = 146.7$ observations. With these parameters, we achieved a total squared deviation between the 20 predictions and observations of 0.0141, which is equivalent to an R^2 of .958. With a mean square error of the means of .0011, this is not a significant difference, $F(15, 345) = 0.89$. Figure 9 illustrates the fit of the model.

Rather than compare this with a specific alternative model, we calculated an upper bound on how well any model could do that did not calculate a correlation and adjust this correlation with a prior bias. Included in this class of models would be the original rational model and any exemplar model that did not incorporate some notion of correlation. Because of the symmetry in the design of the material, any such model must predict that the positive and negative correlations will be exactly inverse. Thus, we used the average of the ith positive and $(6 - i)$th negative quintiles as the predictions for these values separately for defining and nondefining dimensions. Thus, we used the mean of these two numbers, which should be the same under a nonbiased model, to predict the numbers. This can be viewed as a 10-parameter model (five means for defining and five for nondefining dimensions). The correlation between this model and the data was .881. The basic reason for the greater misfit is that the model cannot predict the differential slope for positive and negative correlations in the nondefining dimensions. For instance, in an exemplar model these stimuli have the same number of similar study stimuli the same distance away and so should yield equal estimates of the slope. The modified rational model can fit these different slopes because it incorporates a prior of a positive correlation. An exemplar model might be modified to accommodate this data if it adjusted its predictions by some sort of prior bias about an overall correlation.

Conclusions

This experiment provides further evidence that participants are sensitive to within-category correlation. By taking advantage of the within-category correlation, participants were able to gain accuracy in prediction on dimensions when there was no overall correlation and no difference between the means of the categories. This experiment again showed little difference between participants making up their own categories and participants trained with experimenter categories. Again, this makes sense if participants use within-category correlations. Then their predictions would be much less sensitive to the categories they adopt.

Experiment 3

This experiment was designed to produce even stronger evidence for use of within-category correlation and as strong evidence as is possible against the general category of exemplar models. The basic nature of exemplar models is that they use similar examples to predict properties of a test instance, and what they predict is that the test instance should have similar values. Thus, they will have difficulties making predictions about unseen regions of the instance space, whereas

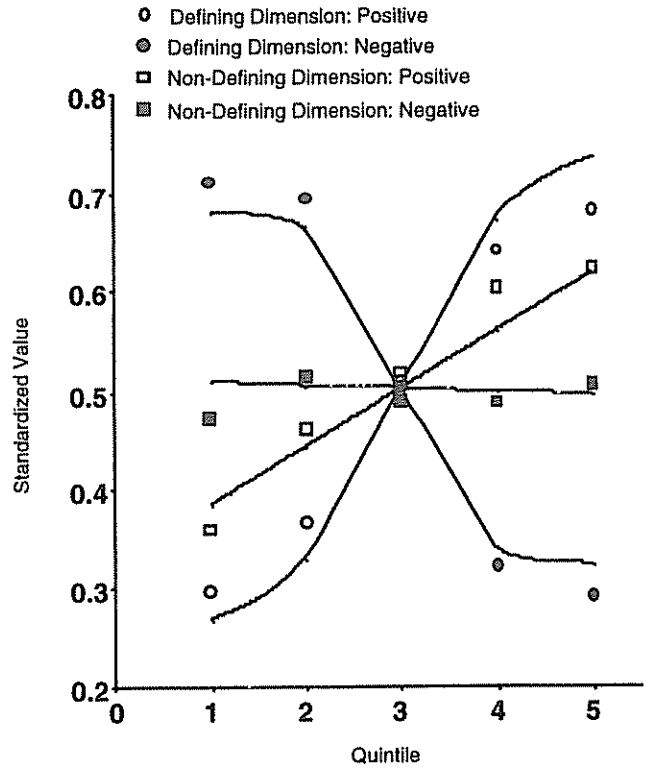

Figure 9 Predictions of the rational model (lines) compared to data (points) for the condition of experimenter categories.

regression models can make such predictions by extrapolating trends. Therefore, we decided to do an experiment using the same basic stimulus structure as the previous experiment, but in which participants would only study instances from the left-hand sides of the sample spaces in Figure 7b but would be tested with both instances from the left-hand and right-hand sides of the space. Thus, they would only see half of the x dimension but would see all of the y dimension. Consider the category with the negative correlation in Figure 7b and suppose the participant is asked to predict an x value given a small y value. According to the correlation, they should predict a large x value, despite having only experienced small x values. However, by any weighting of the experienced values they should predict small x values.

In this experiment we decided to change another aspect of the experiment to avoid participants' bias for positive correlations in the previous experiment. In this experiment the defining dimensions were the size and color of the sepal, and the nondefining dimensions were the size and color of the petal. There was no reason for participants to have any expectations about the sign of these correlations. Finally, to have a reference point to assess the effect of the reduced study space, we tested some of the participants with the full space.

Method

Participants. Fifty-three Carnegie Mellon University undergraduates were recruited to participate in this experiment. The experiment lasted between 90 and 120 min. Participants were paid a flat rate of $6.00. In addition, they received between $3.00 and $6.00 bonus pay that was dependent on performance. Sixteen participants were assigned to the full-range condition, and the remaining 37 participants were in the part-range group and studied part of the space of stimuli.

Materials. The stimuli were generated from the same range of values on the same dimensions as in the previous experiment. There were two categories of flowers. Each category contained 72 instances, yielding a total of 144 flowers. The categories were defined by the values of the sepal features. There were two sepal configurations (counterbalanced across participants). In one configuration, smaller sepals (size = [0, 0.5]) that were light (shade =[0, 0.5]) determined one category, and larger sepals (size = [0.51, 1.0]) that were dark (shade = [0.51, 1.0]) determined the other category. Similarly, the alternative configuration consisted of a category determined by smaller, dark sepals and another category was determined by larger, light sepals. The sepals were constructed such that within a category there was no correlation between sepal size and sepal shade. Note, however, that across the space of sepals from both categories, there was an overall correlation between sepal size and sepal shade.

In the full-range condition, petals were constructed such that the petals in one category exhibited the same correlation as the sepal correlation. That is to say, the within-category correlation of the petals matched the between-category correlation of the sepals. Thus, if the categories were small light sepals and large dark sepals, the petals in the congruent category were such that small went with light and large went with dark. Petals in the other category exhibited the opposite or incongruent correlation. Petals from a particular category were randomly paired with corresponding sepals such that there were no correlations between petal features and sepal features. We counterbalanced across participants how petal correlations were assigned to sepal categories. Because there were two ways of creating sepal categories and two ways of creating petal correlations, there were four conditions of stimulus construction counterbalanced across participants. There were 4 participants in each of the four conditions. Table 3 summarizes the design and category pairings for the full-range group.

Within each category, 36 of the 72 stimuli were randomly selected for the training phase of the experiment. Half of these from each category were randomly chosen to appear in the first training pass, and the remaining half were used for the second training pass. Thus, there were 36 (18 + 18) assigned to the first training pass and 36 (18 + 18) assigned to the second training pass. During the testing phase, all 72 stimuli from each category were presented, for a total of 144 trials. One quarter of these trials involved participants predicting each of the four missing dimensions.

Stimuli for the part-range group were initially constructed by the same procedure as for the full-range group. However, an additional manipulation was performed. During the training portion of the task, we wanted to expose participants to only half of the range of one dimension of the petals. To accomplish this, stimuli were chosen for study subject to the constraint that they had to have values inside the allowable range of the restricted dimension. There were four possible constraints: only light petals (shade = [0, 0.5]), only dark petals (shade = [0.51, 1.0]), only small petals (size = [0, 0.5]), or only large petals (size = [0.51, 1.0]). The constraints were crossed with the category structures in Table 3, yielding a total of 16 (4 × 4) cells in the experiment.

Figure 10 is a graphical representation of selected stimuli subject to the constraint of only small petals: The darker lines represent the correlated petal features that participants would be exposed to during the training phase. There were 36 stimuli from each category that were selected for the training phase. As before, half of these from each category were randomly chosen to appear in the first training pass, and the remaining half were used for the second training pass. Thus, there were 36 (18 + 18) assigned to the first training pass and 36 (18 + 18) assigned to the second training pass. During the testing phase, all 144 stimuli would be presented (72 from each category).

Procedure. The participants were trained to categorize the stimuli into one of two categories according to the same experimenter-defined procedure as that used in Experiment 2. To review, they participated in at least two training passes. During each pass, they were presented with stimuli for which they had to specify the correct category. Feedback was given after each trial. After they correctly categorized

Table 3
Category Structure for Different Groups of Participants

Group	Category 1	Category 2
1	Small light sepals with congruent petal correlation (i.e., small light petals and large dark petals)	Large dark sepals with incongruent petal correlation (i.e., small dark petals and large light petals)
2	Small light sepals with incongruent petal correlation (i.e., small dark petals and large light petals)	Large dark sepals with congruent petal correlation (i.e., small light petals and large dark petals)
3	Small dark sepals with congruent petal correlation (i.e., small dark petals and large light petals)	Large light sepals with incongruent petal correlation (i.e., small light petals and large dark petals)
4	Small dark sepals with incongruent petal correlation (i.e., small light petals and large dark petals)	Large light sepals with congruent petal correlation (i.e., small dark petals and large light petals)

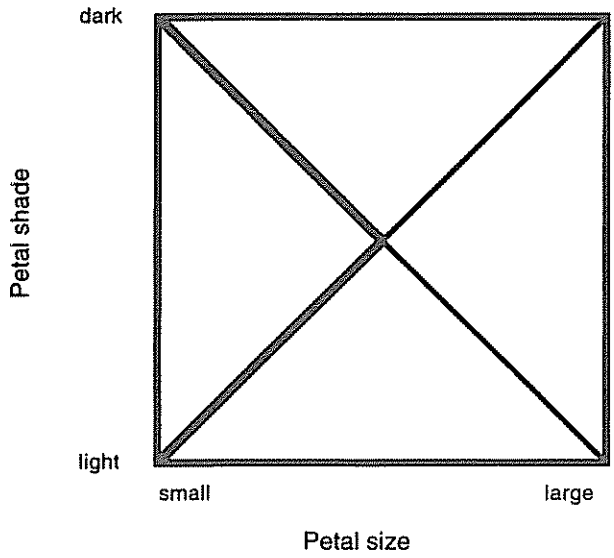

Figure 10. Illustration of stimulus structures for the nondefining dimensions in one part-range condition of the experiment. Bold lines represent stimulus range presented. Bold and light lines represent stimulus range tested.

over 80% of the stimuli in both passes, they could then continue on to the testing phase of the experiment. As in Experiment 2, after selecting the category for a stimulus, participants were presented with a new window that contained a prototype iris (that is, an iris that had the overall median value on all dimensions). They were required to adjust the value on each dimension in order to reproduce the flower they had categorized immediately before. Participants were instructed to perform this task well in order to maximize their accuracy. A point score was displayed at the top of the screen. The greater the accuracy of the reproduction, the more the point score was increased. Participants were told that the final point score would be used to determine to number of bonus dollars they would receive upon completion of the experiment.

After completing the training phase of the experiment, participants were transferred to the prediction phase. They were presented with three of the dimensions and had to predict the fourth. They were tested with all 144 stimuli divided into eight conditions defined by category and dimension to be predicted.

Results

The terms we use in analyzing the results in this section are defined in the Appendix. We refer to the *target* dimension and the *cue* dimension. The target dimension is always the dimension being reproduced. The cue dimension is the one dimension that has a correlation with the target dimension (either within or between categories). It offers the basis for predicting the target dimension. Thus, when predicting a petal dimension, the cue dimension is the other petal dimension, and when predicting a sepal dimension, the cue dimension is the other sepal dimension.

We separately analyzed the data from the defining sepal and nondefining petal dimensions. As in Experiment 2, we converted the data into standardized scores and aggregated the data by the quintile of the cue dimension. Let us consider first the analysis of the defining sepal dimensions. We wanted to assign the cue dimension to quintiles in such a way that we would unify the condition where small and light sepals go together with the condition where small and dark sepals go together. We did this such that the first quintile was always associated with the smallest sepals and whatever shading was associated with small sepals. Thus, if small white sepals were the category and the participant was predicting size (and sepal shading was the cue dimension), the lightest sepals would be in the first quintile and the darkest would be in the fifth. If the participant was predicting shading (and the sepal size was the cue dimension), the smallest sepals would be in the first quintile and the largest would be in the fifth. If one category was small black sepals, the shading assignments were reversed and size assignments maintained. We also adjusted our scoring of participants' shading responses such that dark was considered a small value if they encountered small dark sepals and a large value if they encountered large dark sepals.

An ANOVA was performed on the data from the defining sepal dimensions under this classification. The variables were quintile, whether shading or size was the target dimension, whether participants were exposed to the full range of the nondefining petal dimensions, and the choice for the defining categories. Figure 11 shows the effects of quintile and exposure conditions. The only significant main effect was quintile, $F(4, 196) = 53.64$, $p < .001$, $MSE = 0.013$, and there were significant interactions of this variable with category definition, $F(4, 196) = 10.62$, $p < .001$, $MSE = 0.014$, and with the target dimension, $F(4, 196) = 3.24$, $p < .05$, $MSE = 0.012$. The interaction with category was such that the effects of quintile were steeper when the categories were large white and small black sepals rather than large black and small white sepals (.293 vs. .119 standardized units comparing Quintiles 1 and 5). This may indicate some bias to associate white with large. The interaction with dimension was such that the effects of quintile were steeper when size was the target dimension and shading

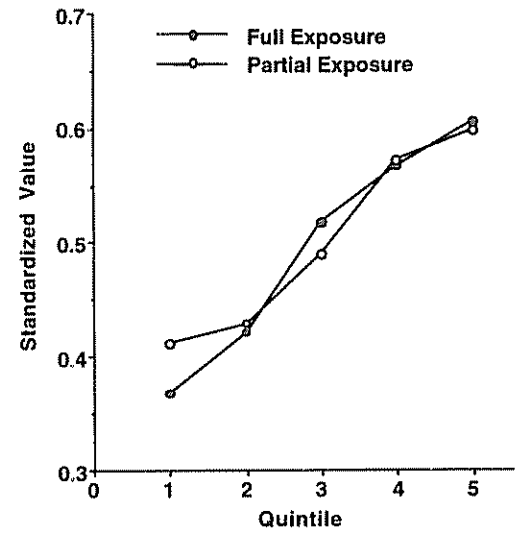

Figure 11. Reproduction of the defining dimension as a function of the value of the cue dimension and exposure to the nondefining dimensions.

was the cue dimension rather than the reverse (.241 vs. .163 standardized units). This may have been due to the fact that it seemed easier to reproduce size. However, basically the results replicate the findings of the previous experiment and further show that with respect to the defining dimensions there was no effect of whether participants experienced the full range of the nondefining dimensions or not.

In the case of the nondefining petal dimensions, we also redefined directions of the dimensions to unify the stimulus conditions. We achieved this by assigning the first quintile to values experienced in that category with the result that the responses involving the extrapolation of the correlation always appeared in Quintiles 4 and 5. Thus, if a participant were experiencing only large dark petals in one category, the first quintile for stimuli in that category would be the darkest shading if shading were the cue and the largest petals if size were the cue. Similarly, we scored the responses such that the large petals and dark petals would be assigned low values. This meant that in all cases a sensitivity to the correlation among the dimensions would show up as a positive relationship between quintile and standardized values of the reproduced scores. Note that when participants experienced the full range of the nondefining dimension, the assignment of low values to petal size and shading was arbitrary and depended on what partial condition they would have been assigned to had they been in the partial condition.

An ANOVA was done on the nondefining petal dimensions where the variables were (a) whether participants experienced the full range of stimuli or only part, (b) quintile, (c) whether the target dimension or the cue dimension had been exposed the full range or not (an arbitrary variable in the case where they saw the full range of stimuli), and (d) whether the relationship between size and shading for that category corresponded to the relationship in the case of the defining dimensions (congruent vs. incongruent). Figure 12 displays the effects of quintile, exposure condition, and congruency. There was a highly significant effect of whether participants had been exposed to the full range of stimuli, $F(1, 51) = 53.95, p < .001$, $MSE = 0.092$). Participants gave much lower standardized values (0.358 versus 0.507) when they had not experienced the full range. The mean of the values they had experienced in the category was 0.25 in the case of having experienced a part range, and it was 0.50 in the case of the full range. Thus, their average values were close to experienced average in the case of the full-range condition but were much larger than what they experienced in that category in the part-range condition. In the part-range condition, they experienced the full range of one dimension across the two categories but only part of the range of the other dimension across both categories. There was a slight tendency for them to reproduce larger values for the dimension for which they had seen the full range of the target dimension across the two categories (0.372 vs. 0.344). This showed up as a significant interaction between group (full range vs. part range) and whether the cue dimension or the target dimension had part range (an arbitrary variable in the full-range condition), $F(1, 51) = 4.95, p < .05, MSE = 0.068$. It is important to remember that although participants experienced the full range of one dimension in the part-range condition, this was across categories, and they only experienced

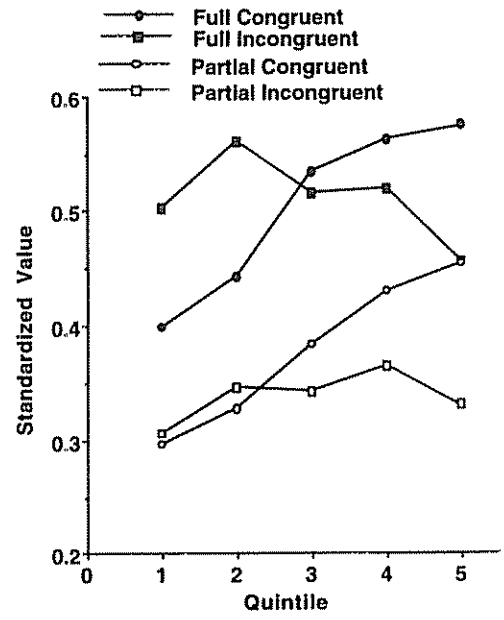

Figure 12 Reproduction of the nondefining dimension as a function of the value of the cue dimension, range of exposure, and congruency with defining dimensions

part of the range within a category. Thus, participants' reproductions for this dimension in the part-range condition were a compromise between the average value experienced within the category (0.25) and the average value experienced for both categories (0.50). However, their reproductions for the other dimension were higher than the mean experienced values in both categories.

The effect of quintile was significant, $F(1, 204) = 10.57, p < .001$, $MSE = 0.022$, and there was a significant interaction of this variable with whether the correlation was congruent with what participants had experienced for the defining sepal dimensions, $F(1, 204) = 10.36, p < .001$, $MSE = 0.027$. The effect of quintiles was large in the case of a congruent relationship (0.165 standardized units comparing Quintiles 1 and 5) but was essentially nonexistent in the case of an incongruent relationship (0.002 standardized units).

Model Fitting

For purposes of fitting models to the data, we decided to display the data separately for the two exposure conditions. Part a of Figure 13 shows the data in the condition where participants were exposed to a full range of stimuli. It shows large effects of quintile both in the case of the defining dimension and in the case of a nondefining dimension congruent with the defining relationship but shows no systematic effect of quintile in the case of the incongruent nondefining dimension. All of these cases are centered around 0.5, which was the mean of the experienced values. Part a of Figure 14 shows the case where participants only experienced a subset of the range of the stimuli. Again there was a large effect of quintile for the defining dimension centered around 0.5. For the nondefining dimensions however, the values were below 0.5. They tended

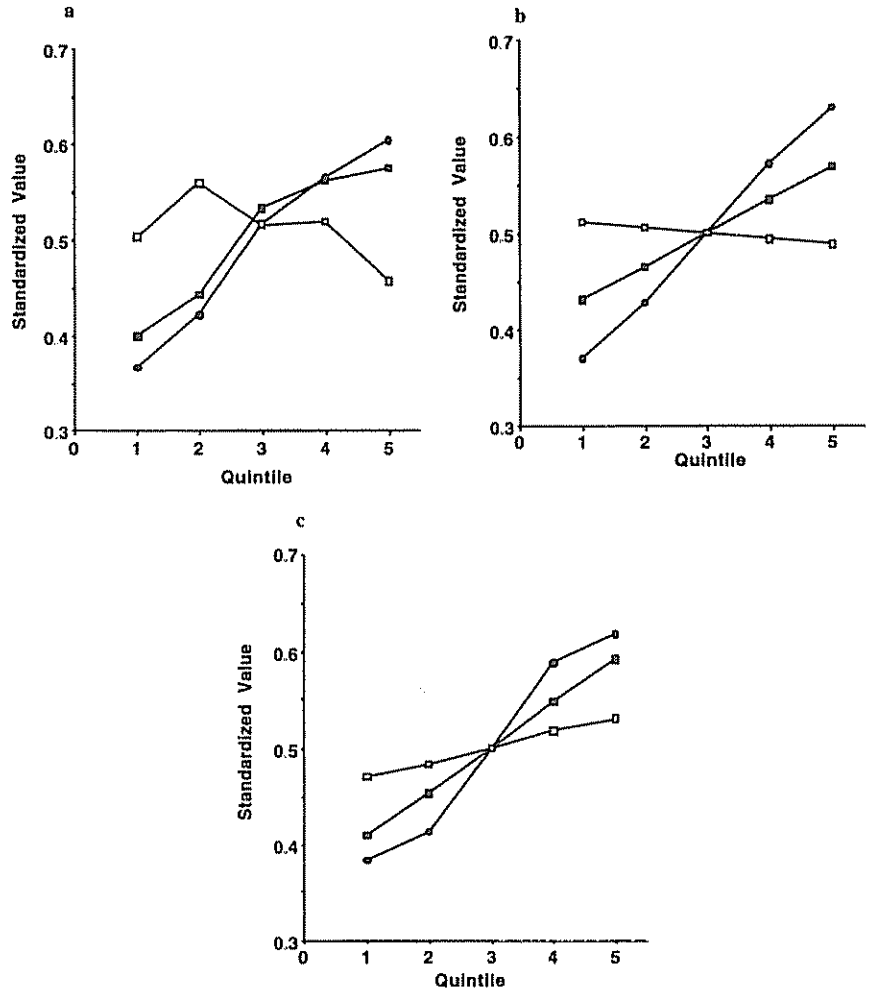

Figure 13 Reproduction of the dimension when participants had experienced the full range of the nondefining dimensions. Part a is participant data, Part b is predictions of the rational model, and Part c is predictions of the exemplar model (with correlations). Data are plotted separately for the defining dimension, for the nondefining dimension for the category that is congruent with the defining dimension, and for the incongruent nondefining dimension. Lines with filled circles represent the defining dimension, lines with filled squares represent the congruent correlation, and lines with open squares represent the incongruent correlation.

to be somewhat lower in the case where the participant experienced the full range of the cue dimension (triangles in Figure 14) rather than the target dimension (squares in Figure 14). For both the cue and the target dimension, there was the same interaction with congruency that is shown in Figure 13a.

Figures 13b and 14b show the outcome of trying to fit the rational model to the data. The same model was used as for Experiment 2, except that rather than using the natural correlation (i.e., large sepals with large petals and dark sepals with dark petals) as the prior, we used the correlation exemplified in the defining dimensions. This can be taken as reflecting a generalization from one pair of dimensions to another. We constrained the value of μ to be 0.5 because assignment of values to end of scale was largely arbitrary. The values of the estimated parameters were $\sigma^2 = 0.520$, $\lambda = 8.24$ observations, $\pi = 0.078$, and $\lambda_\pi = 241.39$ observations. These are similar values to those in the previous experiment. The large value of λ_π combined with a small value of π corresponds to a strong belief that the petal correlation will be weakly in the same direction as the sepals. There are 40 observations represented in the two figures, and with four parameters, the chi-square measure[8] of goodness of fit was 61.80 with 36 degrees of freedom, which, although good, indicates some residual problem. The value of R^2 was .922. A point of discrepancy is that the model overpredicted the mean difference between the cue and target curves (triangles vs. squares),

[8] Because different cells had different standard errors of their means, we calculated $\Sigma_i(x_i - \bar{x}_i)^2/S^2_{\bar{x}_i}$, which is a chi-square statistic rather than the F statistic of the previous experiments. An F statistic is only appropriate when it is reasonable to assume that all cells have similar standard error.

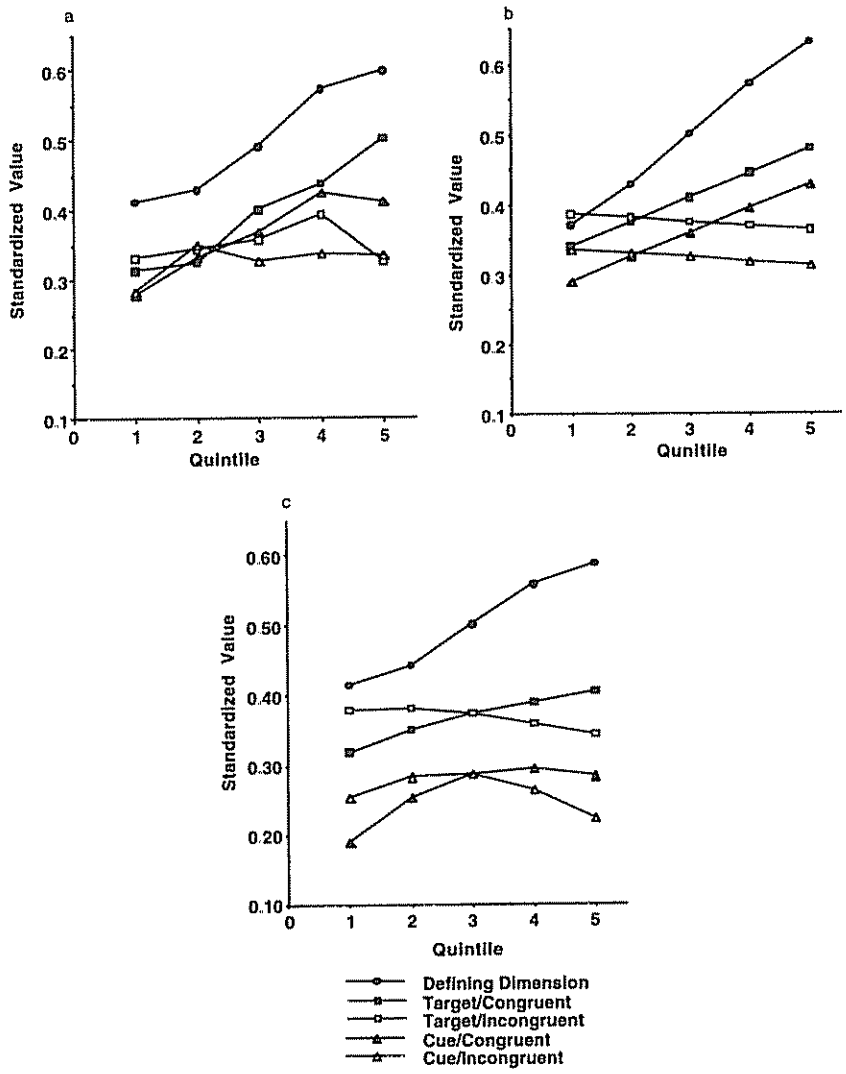

Figure 14. Reproduction of the dimension when participants had only experienced part of the range of the nondefining dimension. Part a is participant data, Part b is predictions of the rational model, and Part c is predictions of the exemplar model (with correlation). Data are plotted separately for the defining dimension, for the nondefining dimension for the category that is congruent with the defining dimension, and for the incongruent nondefining dimension. Also, the data for the nondefining dimension are further divided according to whether the participant experienced the full range of the cueing dimension or the target dimension.

as shown in Figure 14. It predicted a mean difference of 0.052, but the observed difference was only 0.028.

We also tried to fit an exemplar model to the data. We assumed that the participants had exposure in each category to a stimulus set defined by crossing nine values for the defining dimensions with four for the nondefining dimensions ($2 \times 9 \times 4 = 72$ stimuli). The defining values for one category were created by crossing 1/12, 3/12, and 5/12 for each dimension, and the values for the other category were created by crossing 7/12, 9/12, and 11/12. In the case of the nondefining dimensions, we used value pairs of (1/8, 1/8), (3/8, 3/8), (5/8, 5/8), and (7/8, 7/8) for the congruent category and values of (1/8, 7/8), (3/8, 5/8), (5/8, 3/8), and (7/8, 1/8) for the incongruent category. In the part-range condition, participants only had exposure to the first two of these correlated pairs. We then had the model predict the average stimulus for each quintile. We estimated three parameters to reflect the weighting given to the defining dimension, the nondefining dimension with full range, and the nondefining dimension with partial range. These values were estimated to be 1.66, 2.37, and 1.44. The chi-square measure of goodness of fit was 287.95, which is quite large, and $R^2 = .761$. One problem with this model is that it predicts no effect of congruency. We tried a variation on the model in which the prediction was modified by an amount to reflect the effect of congruency. In the congruent case, if the value on the cue dimension was v, then $\alpha(v - 0.5)$ was added to the prediction, whereas this

amount was subtracted in the incongruent case. The weighting parameters estimated for this model were identical, but a value of α was estimated to be 0.077. The chi-square value in this case was 262.77, which can be compared with the rational model that had as many degrees of freedom. The R^2 was .812. Figures 13c and 14c illustrate the fit of this model. One major discrepancy of the model is that it substantially underpredicted the values produced when the participant only experienced part of the target dimension. The mean value experienced on the target dimension was 0.25, and it was difficult for any weighting scheme to produce a mean value different than 0.25. However, the mean value produced by participants was 0.35. The exemplar model failed to predict that the linear relation would continue to increase for Quintiles 4 and 5, where there were no experienced exemplars.

General Discussion

This research started with an attempt to look at human categorization with a set of stimuli that reflected a naturally varying dimensional structure. Starting with stimuli derived from nature rather than testing a theory led us to some rather unexplored ground in the area of categorization and some novel results. Although the second and third experiments went to "controlled" stimuli based on the stimuli of the first experiment, they confirmed and extended the pattern of results. In a nutshell, these experiments provide evidence that participants do create internal categories, do use these to make predictions, and are sensitive to within-category correlations. Unlike in previous research, this sensitivity to correlation cannot be attributed to remembering specific instances or to breaking official categories down into subcategories. However, it needs to be acknowledged that Experiments 2 and 3, which provided the strongest evidence for such correlational sensitivity, had a feature unlike many other experiments. This is that participants were required to reproduce all of the dimensions during training. It is unclear whether such sensitivity would be observed in a paradigm that did not force participants to encode each dimension.

In Experiment 1 we found that participants were much more accurate in their predictions when their prior exposure task caused them to form categories. They were also more sensitive to the within-category, naturally defined dimensional covariation than the rational theory would predict without monitoring of within-category correlation. Also, both Experiments 1 and 2 showed that whether participants identified "true" categories or not had little effect on their predictions. The ability to still predict well given a faulty categorization can be explained if participants were monitoring within-category correlations.

Experiments 2 and 3 showed that participants were sensitive to different correlations within categories. It also showed that they were biased in their expectations about the nature of that correlation. In Experiment 2 these biases came from prior sense of what the "natural correlation" was, but in Experiment 3 participants behaved as if they expected the within-category petal correlation to be the same as the between-category sepal correlation. However, participants were clearly sensitive to the actual within-category correlation in addition to these biases. The sensitivity to different within-category correlations cannot be predicted by category-based models without representation of correlations.

Each experiment found effects of prior bias that were not compatible with published versions of exemplar models. However, the effect in Experiment 1 could be accommodated by adding a prior bias about means, and the effects of Experiments 2 and 3 could be accommodated by adding a prior bias about a sign of an overall correlation. However, it is not so easy to amend exemplar models to deal with the abilities of participants in Experiment 3 to extrapolate category-specific correlations to unexperienced regions of the dimensional space. The most striking disconfirmation of the exemplar model could have come from the part-range condition of Experiment 3 had participants generated values outside of the range they experienced. Although some participants do, this is not true of the average data in Figure 14a. In the data plotted for the cue condition in Figure 14a, participants only generated values of about 0.40 for the highest quintile. However, Figure 14c shows that the exemplar model has great difficulty in accounting even for this high a value. Because these stimuli are so unlike any that the participant has seen, the exemplar model falls back to predicting the overall experienced mean of 0.25. In contrast, the rational model has no difficulty predicting values in the range of 0.40, because its predictions are based on a weighting of the observed strong correlation and a prior weak correlation.

Although we would not want to make the impossible claim that these data disprove any class of theory, we do think we have made the case that they are challenging for certain versions of various classes of theory. They indicate that when participants classify, they are acquiring implicitly a powerful basis for prediction and that this prediction capacity seems sensitive to both the categorical structure of the stimuli and to the within-category correlations.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Anderson, J. R., & Matessa, M. (1992). Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning, 9,* 275–308.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4,* 127–155.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). A Bayesian classification system. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 54–64). San Mateo, CA: Morgan Kaufmann.

Dana, Mrs. W. S. (1893). *How to know the wildflowers*. New York: Scribner.

Elliott, S. W., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 815–836.

Fisher, R. A. (1936). Multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188.

Hammond, K. R., McClelland, G. H., & Mumpower, J. (1980). *Human*

judgment and decision making: Theories, methods, and procedures. New York: Praeger.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 317–330.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin Company.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Mathew, B. (1981). *The iris*. New York: Oxford.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27,* 148–193.

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

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2,* 416–421.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

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

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 908–923.

Appendix

Terms Used in Describing the Results of Experiment 3

Defining dimensions. The two sepal dimensions of size and shading that defined the two categories. A specific category would only have high or low values on these dimensions.

Nondefining dimensions. The two petal dimensions of size and shading. Each category had all possible values on these dimensions. However, the two dimensions had different signs of correlation in the two categories.

Target dimension. The dimension being reproduced.

Cue dimension. The dimension correlated with the target dimension—the other petal dimension if the target dimension were a petal dimension and the other sepal dimension if the target dimension were a sepal dimension.

Congruent correlation. When the correlation displayed within a category for the nondefining petal dimension is the same as the correlation displayed between categories for the defining sepal dimension.

Incongruent correlation. When the correlation displayed within a category for the nondefining petal dimension is the opposite of the correlation displayed between categories for the defining sepal dimensions.

Quintile. Aggregation of the data by fifth of the cue dimension.

Full-range condition. Participants experience the full range of the nondefining petal dimensions in both categories.

Part-range condition. Participants only experience half the range of the nondefining petal dimensions in both categories.

Received September 3, 1993
Revision received March 22, 1995
Accepted March 30, 1995 ■