Perception & Psychophysics
1999, 61 (6), 1178-1199

On the dominance of unidimensional rules in unsupervised categorization

F. GREGORY ASHBY, SARAH QUELLER, and PATRICIA M. BERRETTY
University of California, Santa Barbara, California

In several experiments, observers tried to categorize stimuli constructed from two separable stimulus dimensions in the absence of any trial-by-trial feedback. In all of the experiments, the observers were told the number of categories (i.e., two), they were told that perfect accuracy was possible, and they were given extensive experience in the task (i.e., 800 trials). When the boundary separating the contrasting categories was unidimensional, the accuracy of all observers improved significantly over blocks (i.e., learning occurred), and all observers eventually responded optimally. When the optimal boundary was diagonal, none of the observers responded optimally. Instead they all used some sort of suboptimal unidimensional rule. In a separate feedback experiment, all observers responded optimally in the diagonal condition. These results contrast with those for supervised category learning; they support the hypothesis that in the absence of feedback, people are constrained to use unidimensional rules.

There are three methodologically distinct types of categorization experiments (see, e.g., Ashby & Maddox, 1998). In a *supervised categorization task*, the observer is told the number of contrasting categories, and feedback is provided after every trial. An *unsupervised categorization task* is the same as a supervised task, except that no feedback is provided. In other words, the observer is told the number of contrasting categories, but feedback is never provided. Finally, in a *free sorting task*, observers are not provided feedback, and they are not told the number of contrasting categories. The typical free sorting instructions are to sort the stimuli into as many categories as one desires. In addition, unsupervised and free sorting tasks usually present the stimuli differently. In a typical free sorting task, all the stimuli can be viewed simultaneously, whereas in unsupervised tasks, the stimuli are often viewed sequentially (as they are in supervised tasks).

The vast majority of published categorization studies have been supervised. When feedback is provided, people can learn complex category structures, including those that require integrating two or more stimulus dimensions in a nonlinear fashion (Ashby & Maddox, 1992; McKinley & Nosofsky, 1995). Category learning without feedback has been examined in relatively few studies, and free sorting tasks have been used in most of these. In striking contrast to the results from supervised studies, the consensus is that people use unidimensional rules in the absence of feedback (e.g., Ahn & Medin, 1992; Imai & Garner, 1965; Medin, Wattenmaker, & Hampson, 1987). In other words, they ignore all but one stimulus dimension. On the attended dimension, they generally set a criterion and then assign all stimuli falling below that value into one category and all stimuli falling above that value into the other category.

Despite these results, little is known about the constraints on category learning that might exist in the absence of feedback. This is because, in addition to experimental differences, frequently there are also substantial differences in the goals of supervised and unsupervised tasks. In supervised categorization studies, it is common to ask whether observers *can* learn the underlying category structures. If they fail to achieve some criterion accuracy level after a sufficient amount of practice, we conclude that they have been unable to learn these structures. In other words, some constraints on their category learning abilities have prevented success in this task. In contrast, in unsupervised tasks, it is most common to ask about the preferences of observers. Typically, little or no practice is given, and the design of the experiment provides no motivation for the observers to try to discover the underlying category structure (if one exists). In such experiments, we learn only what categorization strategies people prefer to use. The fact that John arranges books by height does not mean that he cannot learn to arrange books by some more sophisticated strategy.

If one is interested in constraints on unsupervised category learning, rather than on preferences, then at least four experimental design principles must be followed. First, observers must be given extensive practice with the category structures. Second, some underlying category structure must exist; that is, the exemplars of each contrasting category must form a coherent cluster in stimulus space.[1] For example, consider a free sorting or unsupervised cat-

This research was supported by National Science Foundation Grants DBS92-09411 and SBR95-14427 and by a National Science Foundation Graduate Fellowship. We thank William Lee and Leola Alfonso-Reese for their invaluable help with the design and programming of the experiments reported here and Dorrit Billman, Lee Brooks, Kyunghee Koh, Neil Macmillan, Barbara Malt, and Robert Nosofsky for their helpful comments and suggestions. Correspondence concerning this article should be addressed to F. G. Ashby, Department of Psychology, University of California, Santa Barbara, CA 93106 (e-mail: ashby@psych.ucsb.edu).

egorization task in which the stimuli are the integers 1, 2, 3, 4, 5, 6, and 7. All integers between 1 and 7 are represented, and all are equally frequent. Thus, there is no underlying category structure to discover. Even if the observer were told that there were two categories, there would be no rational strategy for discovering these categories. On the other hand, if the same experiment were repeated with the integers 1, 2, 3, 5, 6, and 7, a rational strategy would be to assign the integers 1, 2, and 3 to one category and the integers 5, 6, and 7 to a different category. Third, the observer should be told that there is an underlying category structure. Finally, the observer should be encouraged to respond as accurately as possible.

Unfortunately, even if these design principles are met, there is no guarantee that the resulting data will provide information about constraints on unsupervised category learning. For example, observers might respond optimally because their preferred response strategy was coincidentally the optimal strategy. A simple condition that guarantees that one is studying constraints on learning rather than preferences is that some learning actually occurs—that is, that accuracy improves during the course of the experiment because of some change in the observer's response strategy. An observer who chooses a response strategy because of preference has no reason to change strategies during the experimental session.

There have been many reports of unsupervised categorization studies (e.g., Ahn & Medin, 1992; Boster & D'Andrade, 1989; Homa & Cultice, 1984; Imai & Garner, 1965; Medin et al., 1987; Regehr & Brooks, 1995; Ross, 1996). However, almost none of these has met even two of these four design criteria, and there have been very few demonstrations of learning in unsupervised categorization studies.[2] As a consequence, much is known about preferred response strategies in unsupervised categorization studies, but almost nothing is known about whether there are limits on what people can learn in the absence of feedback.

In this article, we study constraints on unsupervised category learning. In particular, we focus on the following important unanswered questions. What are the constraints on the kinds of decision rules that people use in unsupervised categorization? Have people used unidimensional rules in previous unsupervised studies because of some preference (e.g., perhaps because unidimensional rules are easy to use), or is this the only strategy that people are able to implement in the absence of feedback? Is category learning without feedback possible? And if so, under what conditions should learning be expected?

Two separate unsupervised categorization studies are described. The stimuli in both studies were lines that varied continuously in length and orientation. In each experimental condition, two categories were formed from widely separated coherent clusters of stimuli. In two conditions, a unidimensional rule separated the categories perfectly. In several other experimental conditions, the best unidimensional rule failed badly. In each of the latter conditions, the categories were linearly separable, but the optimal rule had no simple verbal description. All observers in every experiment viewed 800 exemplars from the two categories. There were three principal results. First, all observers in the unidimensional conditions responded optimally during the last experimental block. Second, accuracy in the unidimensional conditions improved significantly over blocks. Thus, observers in the unidimensional conditions learned without feedback. Third, none of the observers in any nondimensional condition responded optimally. Many observers used unidimensional rules, but in contrast to what has been reported in the literature, some observers used rules of a more complex nature. Even so, in contrast to the optimal strategy, none of the observers in these conditions appeared to integrate information from the two stimulus dimensions.

In the next section of this paper, we will briefly review the empirical literature on unsupervised categorization. In the following three sections, we will present our experimental results. The focus of all four of these sections is atheoretical, in the sense that the goal is to describe the empirical phenomena. Throughout these sections, we will refer to theoretically evocative terms, such as *unidimensional rule*, but we will use these terms only as convenient summary descriptions of the data. For example, when we say that people have used unidimensional rules in some experiment, we mean only that their category responses are nicely partitioned on some single stimulus dimension. Specifically, such a statement is not a claim about a psychological process. Indeed, it is well known that many different categorization theories can account for such "unidimensional" responding (see, e.g., Ashby & Maddox, 1998). In the General Discussion (i.e., the sixth section), we will explicitly consider the theoretical implications of our data. Finally, we will close with some general comments and conclusions.

PREVIOUS CATEGORIZATION STUDIES WITHOUT FEEDBACK

Before we describe the literature on unsupervised categorization, it is important to note that there have been many studies in which no feedback was given and the tasks were other than categorization; in some of these, the tasks have been closely related to categorization. For example, Clapper and Bower (1991, 1994) had observers view exemplars of a category and then list the features that were most informative for distinguishing each specific exemplar from the other category members. As another example, Billman and Knutson (1996) had observers view pictures of imaginary animals. Next, the observers were shown pictures of two new animals and were asked to select the one most consistent with the animals in the training set.

Success in either of these tasks required observers to learn the frequencies of various stimulus features and their interrelationships. Even so, such knowledge does not guarantee an ability to assign stimuli to categories. For example, consider the category structures described by the

Table 1
Category Structure Used by Medin, Wattenmaker, and Hampson (1987)

	Category A Dimension				Category B Dimension			
Stimulus	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
1	1	1	1	1	0	0	0	0
2	1	1	1	0	0	0	0	1
3	1	1	0	1	0	0	1	0
4	1	0	1	1	0	1	0	0
5	0	1	1	1	1	0	0	0

bottom two panels of Figure 2. Each point in these panels describes a different line that varies in length and orientation. The "+" signs indicate the lengths and orientations of the exemplars of one category, and the "○" signs describe the exemplars of a contrasting category. An observer who had no idea that there were two separate categories might still detect the correlation between length and orientation that exists in the overall combined stimulus ensemble (i.e., a positive correlation in the left panel, and a negative correlation in the right panel). For example, in the diagonal–positive condition, such an observer would know that long lines tended to occur only with large orientations, even though he or she would be at chance in categorization. Thus, unsupervised studies in which observers try to learn about stimulus features provide only limited information about unsupervised categorization.

The first major study of categorization behavior in the absence of feedback was done by Imai and Garner (1965). Each of their stimuli was a white card depicting two dots that varied in overall position, interdot distance, or orientation. On each trial, observers were shown either eight or four cards and were asked to sort these into two piles in any way that seemed reasonable. Stimuli on eight-card trials were always constructed by factorially crossing two levels from each of the three stimulus dimensions, and four-card trials were always constructed by factorially crossing two levels from two of the three stimulus dimensions. All observers showed a strong preference for unidimensional rules. With the factorial stimulus design used by Imai and Garner, however, there are no stimulus clusters. Thus, there is no a priori category structure for observers to discover. For this reason, any decision rule chosen by observers is as good as any other rule.[3] Consequently, the results of Imai and Garner tell us that humans prefer unidimensional rules, but they tell us nothing about constraints on human category learning in the absence of feedback or even whether learning without feedback is possible.

Medin et al. (1987) introduced an experimental design that has been used in a number of subsequent unsupervised categorization studies (e.g., Regehr & Brooks, 1995). The stimuli were insect-like creatures that varied on four binary-valued dimensions. As in the Imai and Garner (1965) study, Medin et al.'s observers showed a strong tendency to use unidimensional rules. Table 1 shows the structure of the two categories that Medin et al. used. The separate categories form coherent clusters in the four dimensional binary-valued stimulus space. But consider the perceived category structure for an observer unable to attend to all four stimulus dimensions. For example, suppose an observer is able to attend to only three of the four stimulus dimensions. Examination of Table 1 indicates that no matter which of the dimensions is ignored, the perceived stimuli will have the structure shown in Figure 1. The solid dots in Figure 1 represent perceived stimuli that occur with a probability of .2, and the open dots represent perceived stimuli that occur with a probability of .1. Given that the perceived stimulus structure shown in Figure 1 does not contain separate clusters (see note 1), how could an observer decide which stimuli go in which categories? If it was known that there were two categories, an observer might hypothesize that the two high-frequency stimuli belong to different categories, but the category assignments for the other six stimuli would have to be made on some arbitrary basis.[4] Thus, the fact that observers acting on the Table 1 design invariably used unidimensional rules tells us little about whether they could learn to use more than one stimulus dimension, unless we assume that they were able to attend perfectly to all four stimulus dimensions.

Ahn and Medin (1992) improved on this design by using stimulus dimensions that, although still discrete valued, had more than two possible levels. This increased the number of possible stimuli (i.e., to n^d, where n is the number of levels per dimension and d is the number of dimensions), and so made it more likely that structures like those shown in Table 1 would form coherent clusters in cases in which observers attended only to some subset of possible stimulus dimensions. Even so, the categories used by Ahn and Medin each contained only five exem-

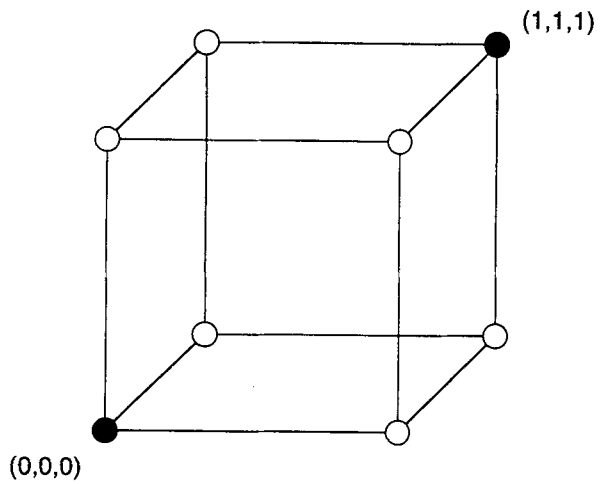

Figure 1. Perceived stimulus structure in the unsupervised categorization experiment of Medin, Wattenmaker, and Hampson (1987; see our Table 1) for an observer who ignores any one of the four stimulus dimensions.

plars. Furthermore, in each case an observer could achieve perfect performance by using a unidimensional rule and memorizing only two exceptions. Ahn and Medin found unidimensional responding in some conditions, but in other conditions the responses of a majority of observers were more consistent with a similarity-based (i.e., nondimensional) rule. Because of the category structure, however, it is difficult to know whether observers in the latter conditions used a true similarity-based rule or used a unidimensional rule and memorized the few exceptions. The latter possibility is hypothesized, for example, by the RULEX model of category learning (Nosofsky, Palmeri, & McKinley, 1994).

Thus, the literature on unsupervised categorization clearly establishes that people typically use unidimensional rules when there is no compelling reason to do otherwise. There are few published data, however, on the question of whether this tendency is a preference or a real constraint on human performance. Another complication is that in almost all of the published studies, discrete-valued rather than continuous-valued stimulus dimensions have been used. Binary-valued dimensions are especially problematic, because they might encourage people to use unidimensional rules in situations in which they might normally use a rule that integrates information across dimensions. For example, in many cases in which binary-valued dimensions are used, the two levels can be interpreted as the presence and absence of a stimulus feature (e.g., *has wings or not*, *has symptom X or not*). Tversky (1972, 1977) argued persuasively that the psychological representation of such features is nonnumeric. If so, it might be especially difficult for people to integrate information across such dimensions (or features)—as is required, for example, to learn the category structures shown in the bottom two panels of Figure 2. Thus, it is important to study unsupervised categorization performance with stimuli constructed from continuous-valued dimensions.

Another important question, which is related to the issue of whether people are constrained to use unidimensional rules, is whether category learning is possible without feedback. Few studies in the literature speak to this issue. For example, each observer in the Ahn and Medin (1992) study categorized a total of only 10 stimuli, and these observers never categorized the same stimulus more than once. Since learning is by definition a change in performance across replications, Ahn and Medin and other such studies say nothing about learning. A few older studies showed modest levels of learning in free sorting tasks (Aiken & Brown, 1971; Evans & Arnoult, 1967), but perhaps the most dramatic examples of unsupervised category learning were reported by Fried and Holyoak (1984) and by Homa and Cultice (1984). The stimuli in both studies were complex visual patterns that varied on many stimulus dimensions (10×10 grids of randomly distributed light and dark squares; lines that connected nine randomly located dots), and in both cases the contrasting categories were created by randomly distorting prototype patterns. Fried and Holyoak reported learning in a variety of unsupervised conditions, whereas Homa and Cultice found no learning across trials in unsupervised conditions in which the category exemplars were created from moderate- or high-level distortions of the prototype. However, when the categories contained only low-level distortions of the prototype, learning was significant (accuracy improved from 68% to 92% across trials).

The results of Fried and Holyoak (1984) and Homa and Cultice (1984) indicate that unsupervised category learning is possible, at least under certain conditions in which category exemplars are created from random distortions of the category prototype. This is an important result, but although the stimuli that were used in these studies were perceptually interesting, it is difficult to draw stronger conclusions. The stimuli used by Fried and Holyoak varied on 100 physical dimensions, and the Homa and Cultice stimuli varied on 18 physical dimensions (i.e., 2 spatial coordinates define each of the 9 dots). In both cases, however, the dimensionality of the psychological representation is unknown. In particular, there is good evidence that the psychological dimensions of these complex stimuli were not based in any simple way on the physical dimensions (Shin & Nosofsky, 1992), and it is not known whether the stimuli used in either study could be sorted successfully by observers using unidimensional rules. Also, without knowing the psychological representation, it is impossible to know how widely separated the category clusters were in psychological space.

In summary, there is overwhelming evidence that humans prefer to use unidimensional rules, at least when the stimulus dimensions are binary or trinary valued. There are even reports that, under certain conditions, people will use nondimensional rules (e.g., Ahn & Medin, 1992), but the design of these studies makes it difficult to rule out augmented unidimensional strategies (e.g., unidimensional rule plus exceptions). There is also evidence that, under certain conditions, learning without feedback is possible, but it is not clear whether humans can learn nondimensional rules without feedback. In addition, there has been no attempt to compare systematically (1) the ability of people to learn unidimensional rules without feedback with (2) the ability of people to learn nondimensional rules without feedback.

EXPERIMENT 1A

In Experiment 1A, we compared the ability of people to learn unidimensional categorization rules with the ability of people to learn nondimensional categorization rules, when the categories formed coherent clusters and the stimulus dimensions were continuous-valued, but when no feedback was provided. To ensure that failures of learning would be due to some constraint on the abili-

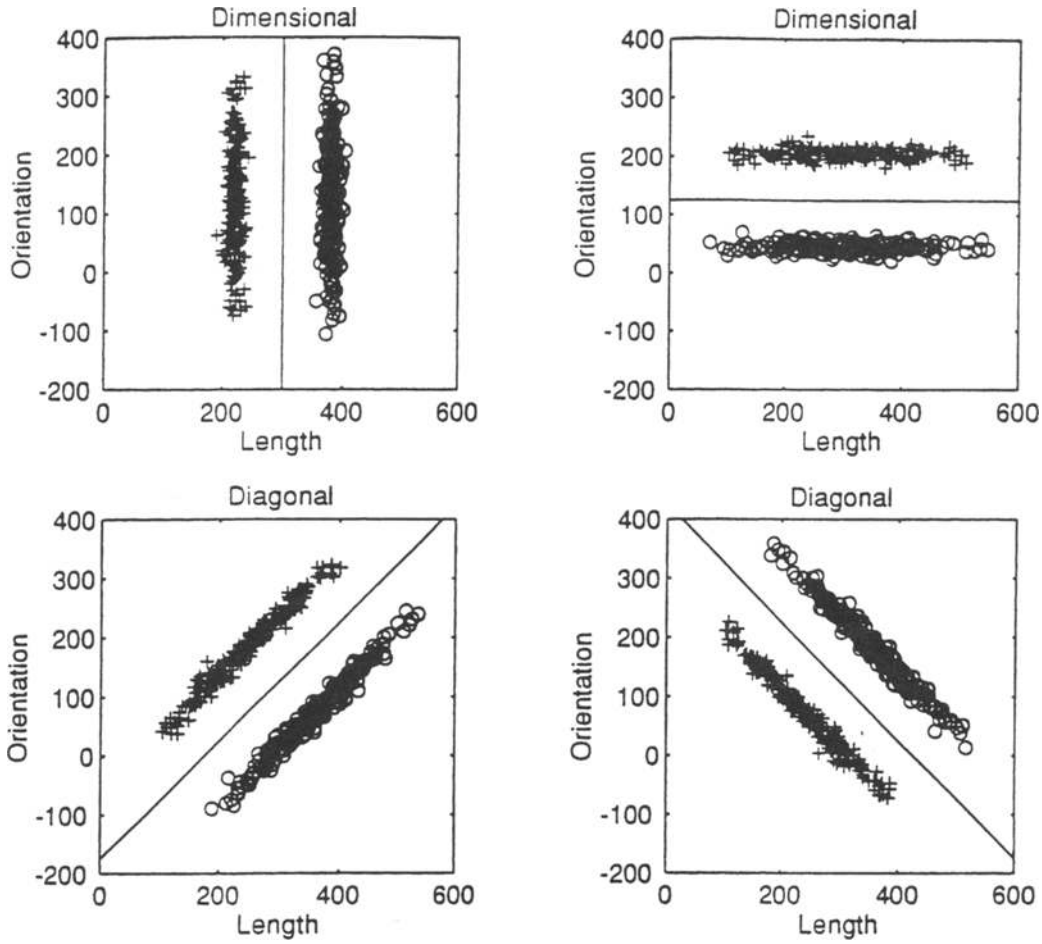

Figure 2. Category structure of the four Experiment 1A conditions. Each plus sign depicts the coordinates of a stimulus in one category, and each circle depicts the coordinates of a stimulus in the contrasting category. Upper left, unidimensional–length condition; upper right, unidimensional–orientation condition; lower left, diagonal–positive condition; lower right, diagonal–negative condition.

ties of the observers, rather than on their preferences, extensive practice was given, and the observers were told that perfect accuracy was possible.

The design of all the experiments described in this article was based on the randomization technique introduced by Ashby and Gott (1988). The stimuli in all experiments were lines that varied continuously in length and orientation. These dimensions are perceptually separable and the transformation from the physical to the perceptual space involves only minor distortions[5] (for discussions of this point, see, e.g., Ashby & Lee, 1991; Ashby & Maddox, 1990; Nosofsky, 1986). The category structures used in Experiment 1A are shown in Figure 2. Each symbol in Figure 2 represents a single stimulus (the plus signs and circles represent stimuli in the two different categories). In each condition, there were two distinct categories that did not overlap, so perfect accuracy was always possible.

The category structures in the four conditions were generated by successively rotating the categories shown in the upper left panel of Figure 2 by increments of 45°. Thus, by any of the objective measures that are popular in cluster analysis (e.g., Fukunaga, 1990), task difficulty was invariant across the four conditions. Specifically, maximum possible accuracy, within-category scatter, between-category separation, and category coherence were all identical in the four conditions. Also shown in Figure 2 are the decision bounds that maximized categorization accuracy. These are the lines $y = x - 175$, $x = 300$, $y = 125$, and $y = -x + 425$. In two conditions, the optimal bound was unidimensional, and in two conditions, it was diagonal. In the two diagonal conditions, the most accurate unidimensional rule yielded a response accuracy of about 80%. In addition, because of the continuous-valued stimulus dimensions, it would have been difficult or impossible to respond optimally in the diagonal conditions by using a unidimensional rule and memorizing all of the exceptions. Figure 3 shows a few exemplars from each category in one unidimensional condition and one diagonal condition.

Experiment 1A was subdivided into separate blocks of 80 trials each. On the odd-numbered blocks, observers passively viewed the 80 stimuli without responding. On the even-numbered blocks, observers assigned each stimulus to category A or B, by pressing an appropriate button.

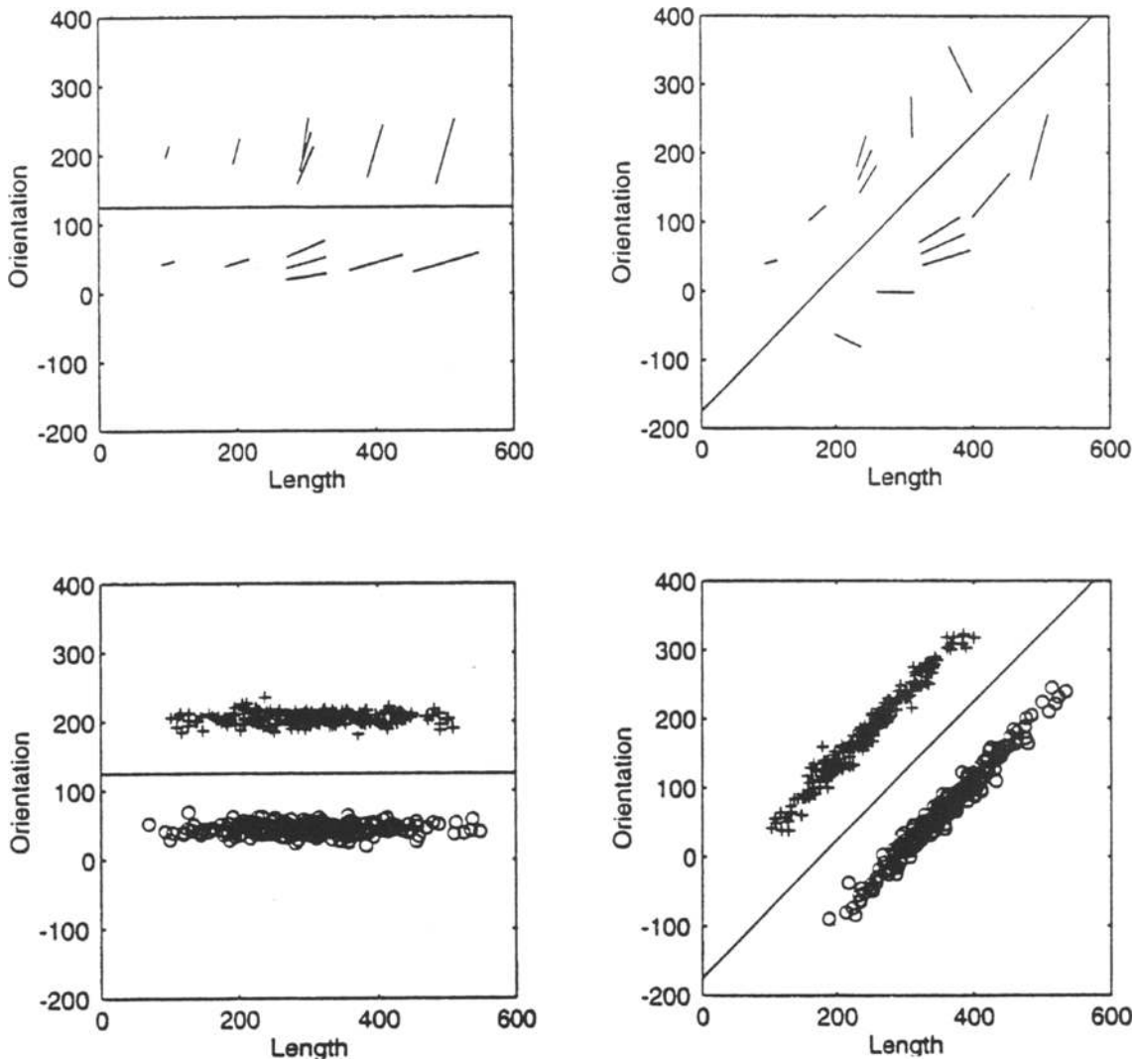

Figure 3. Category structure from the unidimensional–orientation and diagonal–positive conditions of Experiment 1A, together with a few representative stimuli from each of the contrasting categories.

They were told that it made no difference which category they called A, but that it was important to maintain consistency in their definition. They were also told that there were only two categories and that perfect accuracy was possible. During the experiment, however, they were never given feedback about their accuracy, either on single trials or on blocks of trials.

Although it may be obvious, it is important to note that, even if observers are constrained to use unidimensional rules, they will respond optimally in the unidimensional conditions only if they manage to solve two problems. First, they must determine whether length or orientation is the critical dimension; then, they must discover the optimal criterion value along that dimension (i.e., the optimal intercept).

Method

Observers and Design. The observers in all experiments were undergraduates at the University of California at Santa Barbara who received partial course credit in an introductory psychology course for participating in the experiment. Twenty observers participated in Experiment 1A. Five observers were assigned to each of four experimental conditions, which differed according to the orientation of the boundary that separated the two categories. The four conditions are illustrated schematically in Figure 2.

Stimuli and Apparatus. The stimulus sets from the four different experimental conditions are shown in Figure 2. Each point represents the position of a stimulus in the orientation × length space. The category A stimuli are plotted as plus symbols and the category B stimuli are plotted as circles. In each condition, stimuli were generated by randomly sampling from each of two bivariate normal distributions. The two category distributions were specified by a mean and variance on each dimension and a covariance between dimensions. The two categories always had different means but the same variances and covariance. See Table 2 for the exact parameter values.

Using the parameters listed in Table 2 for the unidimensional–length condition, 200 random samples were drawn from the category A population and 200 were sampled from the category B population. A linear transformation was then performed on each sample so that the sample statistics exactly matched the Table 2 population

Table 2
Parameter Values That Define the Categories Used in Experiment 1A

	Means				Variances		
	Length		Orientation				
Condition	Category A	Category B	Category A	Category B	Length	Orientation	Covariance
Unidimensional							
Orientation	300	300	205	45	9,000	75	0
Length	220	380	125	125	75	9,000	0
Diagonal							
Positive	243	357	68	182	4,538	4,538	4,463
Negative	243	357	182	68	4,538	4,538	−4,463

parameters. For each of the 5 observers in the unidimensional–length condition, this sample of 400 stimuli was randomly partitioned into five blocks with 80 stimuli in each block. Each observer was presented with a different random partitioning of these 400 trials. These stimuli constituted the response blocks (i.e., the even-numbered blocks) for each observer. Using these same procedures, a second sample of 400 trials was generated for each observer. These stimuli constituted the observation blocks (i.e., the odd-numbered blocks).

For each observer, the stimuli from the response blocks and the stimuli from the observation blocks were combined in one stimulus set in such a way that the observation blocks constituted the odd-numbered blocks and the response blocks constituted the even-numbered blocks. The stimuli in the other three conditions were generated by rotating the stimulus set from the unidimensional–length condition: (1) 45° in the diagonal–positive condition, (2) 90° in the unidimensional–orientation condition, and (3) −45° in the diagonal–negative condition. The stimulus set for each observer was centered about an orientation of 45° and a line length of 300 pixels. The distribution parameter values for each of the four conditions are listed in Table 2.

The stimuli were computer generated and displayed on a Mitsubishi Electric Color Display Monitor Model C-9918NB in a dimly lit room. Each random sample (x_1, x_2) was converted to a stimulus by letting x_1 determine length of the line and x_2 determine orientation. For example, the category A mean in the unidimensional–length condition was converted to a line 220 pixels long, rotated $125 \times (\pi/500)$ radians counterclockwise from horizontal. The $(\pi/500)$ scaling factor was chosen in an attempt to equalize the salience of orientation and length. The stimuli were presented in white on a dark background, and the visual angle of the stimuli ranged from about 1.5° to about 6.3°.

Procedure. Each observer was run individually. The observers were told that the stimuli would be presented one at a time on a monitor and that their task was to separate the stimuli into two categories of equal size. Five observation-only blocks alternated with five response blocks. During each observation block, the observers were instructed simply to look at 80 sequentially presented stimuli and to try to learn about the two categories. Each stimulus was presented for 1 sec with a 500-msec interstimulus interval. During each response block, the observers were instructed to select a category for each stimulus and to press a button labeled "A" or a button labeled "B" to show which category had been selected. The observers were informed that the category labels were arbitrary and were warned to be consistent about what they called a member of category A and what they called a member of category B. Since category assignment to the labels "A" and "B" was arbitrary, each observer was assumed to have assigned category labels in the manner that resulted in the highest calculated percent correct within each block. With this assumption, it was impossible for observers to score less than 50% correct. The observers were told that perfect accuracy was possible, but they were never given any feedback about their performance. During the response blocks, the stimuli were response terminated (with a 5-sec maximum exposure duration), and the interstimulus interval was 1 sec. The break between blocks was observer paced.

Results and Discussion

Accuracy analysis. Recall that 5 observers were run in each of the four conditions and that by using the optimal decision bound, observers in any of the four conditions could achieve perfect accuracy. Also, recall that each observer was assumed to have assigned category labels in the manner that resulted in the highest calculated percent correct during each response block. Table 3 shows the percent correct for each observer during the last response block. Note that accuracy averages 98% correct in the unidimensional conditions, but only 62.4% correct in the diagonal conditions. Fifteen of the 20 observers performed better in some block other than the last. Recomputing the average percent correct using the best block for each observer does not eliminate the substantial advantage of the unidimensional conditions (99.13% vs. 75.5%). Similarly, overall accuracy was higher for observers in the unidimensional conditions than for observers in the diagonal conditions (91.7% vs. 65.3%). These results indicate that by the end of the session, observers in the unidimensional conditions were responding almost perfectly. In contrast, observers in the diagonal conditions performed poorly throughout the experiment. In fact, the accuracy data provide no evidence that observers in the diagonal conditions ever integrated information across stimulus dimensions. In 49 of the 50 response blocks in the diagonal conditions (i.e., 10 observers × 5 blocks), accuracy

Table 3
Percent Correct for Each Observer During the Last Response Block of Experiment 1A

	Observer				
Condition	1	2	3	4	5
Unidimensional					
Orientation	96.3	100	100	90	97.5
Length	98.8	98.8	98.8	100	100
Diagonal					
Positive	51.3	75	76.3	71.3	62.5
Negative	56.3	66.3	55	53.8	65

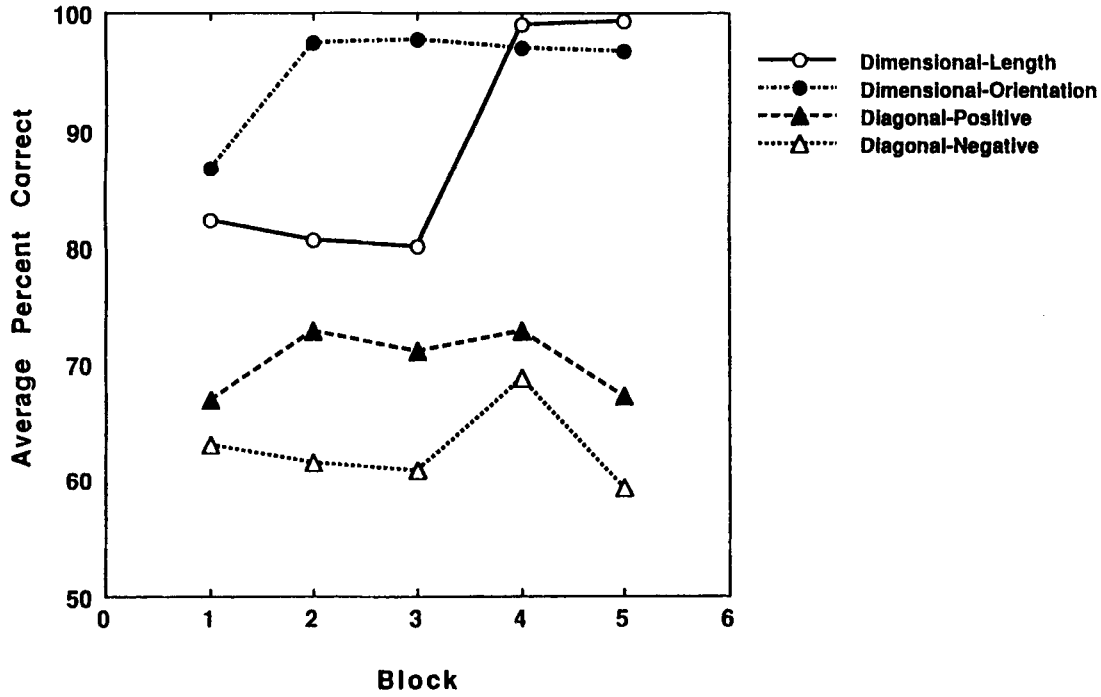

Figure 4. Mean percent correct by block in each of the four conditions of Experiment 1A.

was poorer that one would expect if the observers had used the most accurate unidimensional rule.

Figure 4 shows the average percent correct during each response block in all four conditions. Note that no learning occurred in either diagonal condition; average accuracy was no higher in the last block than in the first. In contrast, in both unidimensional conditions, accuracy increased dramatically during the course of a single block, from about 84% correct to almost perfect accuracy. By definition, such an increase in accuracy implies that learning occurred, and it suggests some sort of change in the response strategies used by observers. Thus, observers in the unidimensional conditions did not persist in using the same rule that they initially preferred. Although all observers in both unidimensional conditions eventually responded with almost perfect accuracy, learning occurred more quickly in the unidimensional–orientation condition than in the unidimensional–length condition (i.e., on Block 2 vs. Block 4). One reason for the quicker learning in the unidimensional–orientation condition might be that the criterion orientation that separated the two categories was close to 45° (see Figure 3). Although there was no natural linguistic marker for this orientation (unlike 0° and 90°), it still might have served as a prelearned reference point. On the other hand, it is important to note that despite this possibility, the performance of observers in the unidimensional–orientation condition *did* improve, which indicates that their eventual optimal performance was not due simply to an initial preference for the optimal decision rule.

Figures 5–8 show the actual responses of each observer during the last experimental block. A plus indicates a response of one type, and a circle indicates the other type of response. For the moment, the lines may be ignored. These figures show clearly the dramatic difference in performance in the unidimensional as opposed to the diagonal conditions. Whereas observers responded almost perfectly in the last block of the unidimensional conditions, the responses in the diagonal conditions show no evidence that any observers detected the underlying category structure.

Model-based analysis. To get a more detailed picture of how observers categorized the stimuli, a number of different models derived from decision bound theory (Ashby, 1992; Maddox & Ashby, 1993) were fit to each observer's responses. Decision bound theory assumes that each observer partitions the perceptual space into response regions by constructing a decision bound. On each trial, the observer determines which region the percept is in and then emits the associated response. Despite this deterministic decision rule, decision bound models predict probabilistic responding because of trial-by-trial perceptual and criterial noise. We fit five different versions of decision bound theory to the data collected in Experiment 1A. All of the models, except for the interval-based unidimensional classifier, are described in detail by Ashby (1992).

The goal of the analyses reported in this section is to obtain the best possible description of the data from each individual observer. For example, this analysis will allow

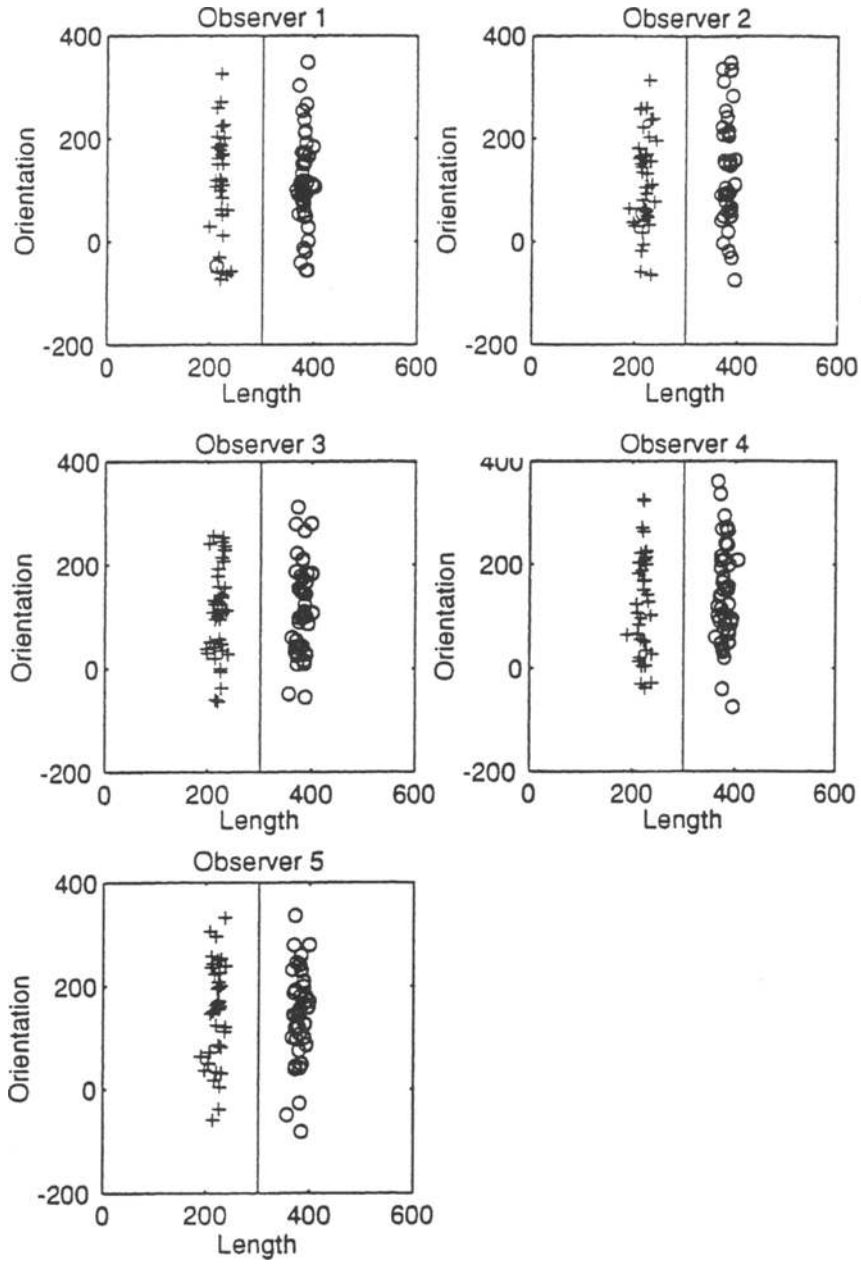

Figure 5. Category responses during the last block for all observers in the unidimensional–length condition of Experiment 1A.

us to determine whether the violations from unidimensional responding for any given observer are statistically significant. It is important to note, however, that a good fit of any specific model provides only limited information about psychological process. In particular, it is likely that some model making very different process assumptions (e.g., an exemplar-based model) might fit as well as the best of these five decision bound models. With this caveat in mind, we proceed with a description of the five decision bound models.

General linear classifier. The general linear classifier (GLC) assumes that the decision bound is linear. Maddox and Ashby (1993) found that the GLC accounted for categorization data about as well as the most powerful exemplar models in experiments in which the optimal decision bound was linear. In the present applications, the GLC has three free parameters: the slope and intercept of the linear decision bound and the variance of internal (perceptual and criterial) noise (i.e., σ^2).

Unidimensional classifiers. The unidimensional classifiers assume that observers use a unidimensional rule (i.e., a vertical or horizontal decision bound). These models each have two free parameters: the intercept of the decision bound and σ^2.

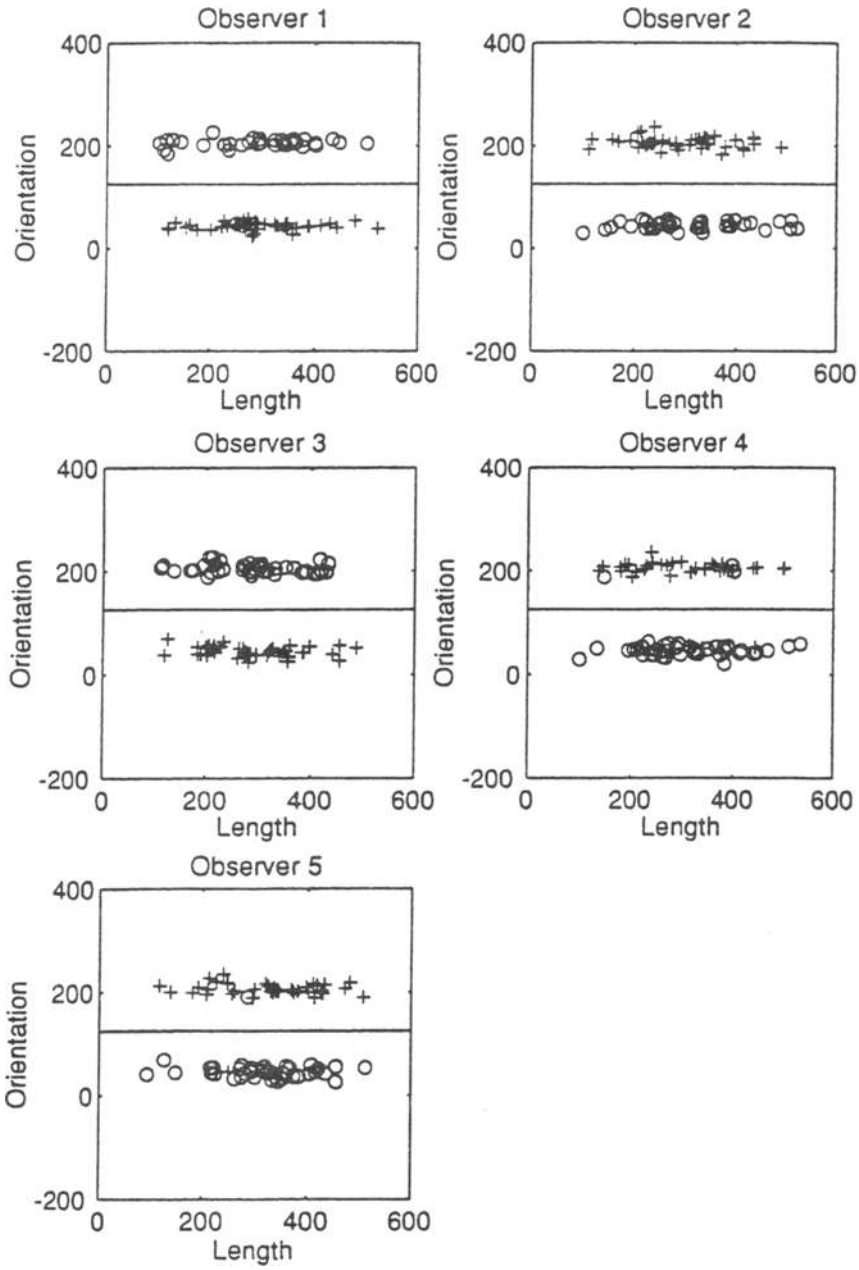

Figure 6. Category responses during the last block for all observers in the unidimensional–orientation condition of Experiment 1A.

Optimal classifier. This model assumes that observers use the decision bound that maximizes accuracy (shown in Figure 2). With the category structures used in Experiment 1A, the optimal decision bounds are all linear. This model has only a single free parameter (i.e., σ^2).

Interval-based unidimensional classifier. Some of the observers in the diagonal conditions appeared to use a generalized unidimensional strategy in which all stimuli with orientations within some interval (e.g., 0°–90°) are assigned to one category and all stimuli with orientations out of this interval are assigned to the contrasting category. To test this hypothesis, we developed an interval-based unidimensional model with two horizontal decision bounds (since the y-axis is orientation) (see also Nosofsky, Clark, & Shin, 1989). This model has three free parameters: the intercepts of the two horizontal bounds and σ^2.

Using an iterative maximum likelihood parameter estimation procedure, each of these models was fit separately to the data from each response block of every observer. Data from separate blocks were fit because it was apparent that observers sometimes switched strategies from

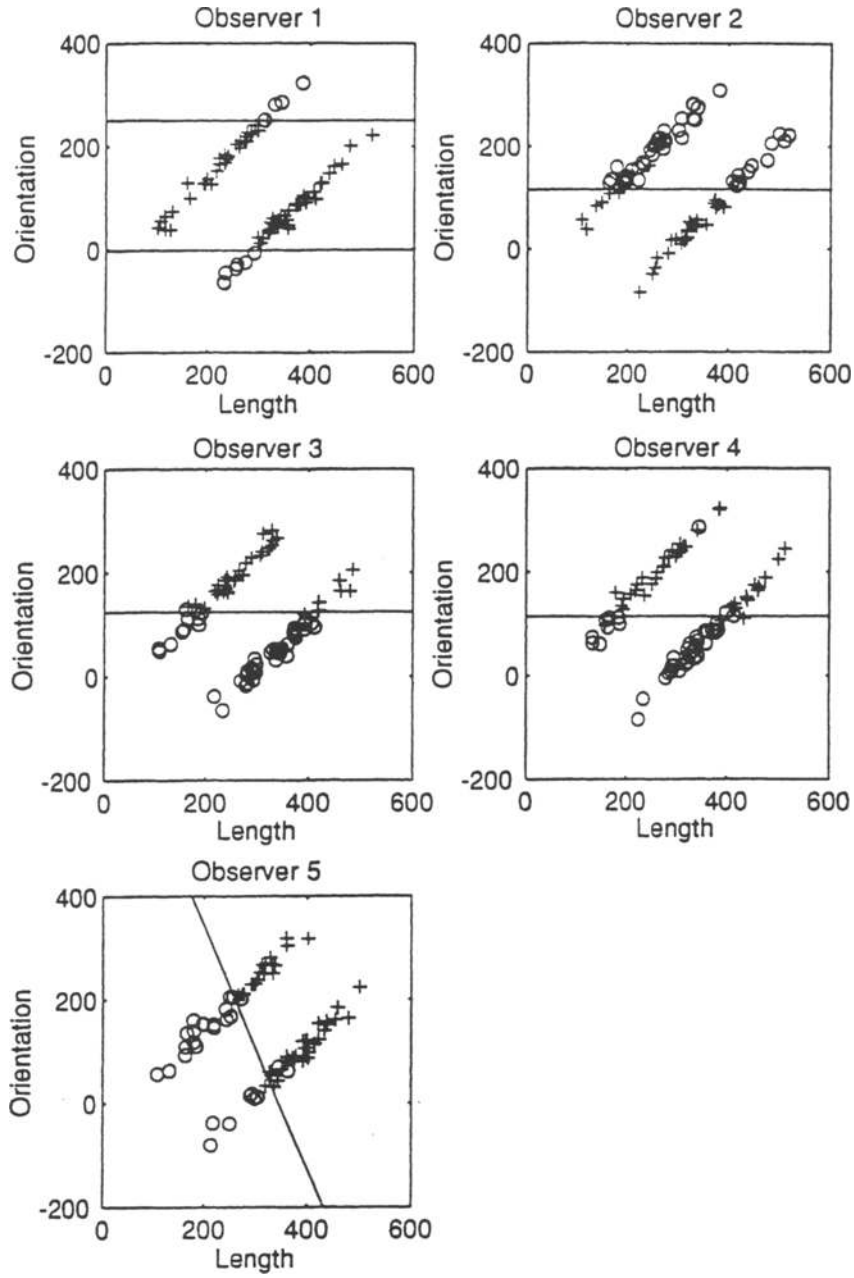

Figure 7. Category responses during the last block for all observers in the diagonal–positive condition of Experiment 1A.

block to block. To select the best-fitting model, we used the A information criterion (AIC) of Akaike (1974; see also Takane & Shibayama, 1992):

$$\text{AIC} = -2L + 2v,$$

where v is the number of free parameters and L is the log likelihood of the data, given the model. The AIC statistic penalizes a model for extra free parameters in such a way that the smaller the AIC, the closer a model is to the "true model," regardless of the number of free parameters. As a result, to find the best model among a given set of competitors, one simply computes an AIC value for each model and chooses the model associated with the smallest AIC.

Table 4 shows the number of times each of the five competing models provided the best fit to the last block of data, and Table 5 shows the number of times each model provided the best fit to any block of data. The decision bounds from the best-fitting model are illustrated for the last block by the solid lines in Figures 5–8. To begin, consider the results from fitting the models to the data collected from the unidimensional conditions. In

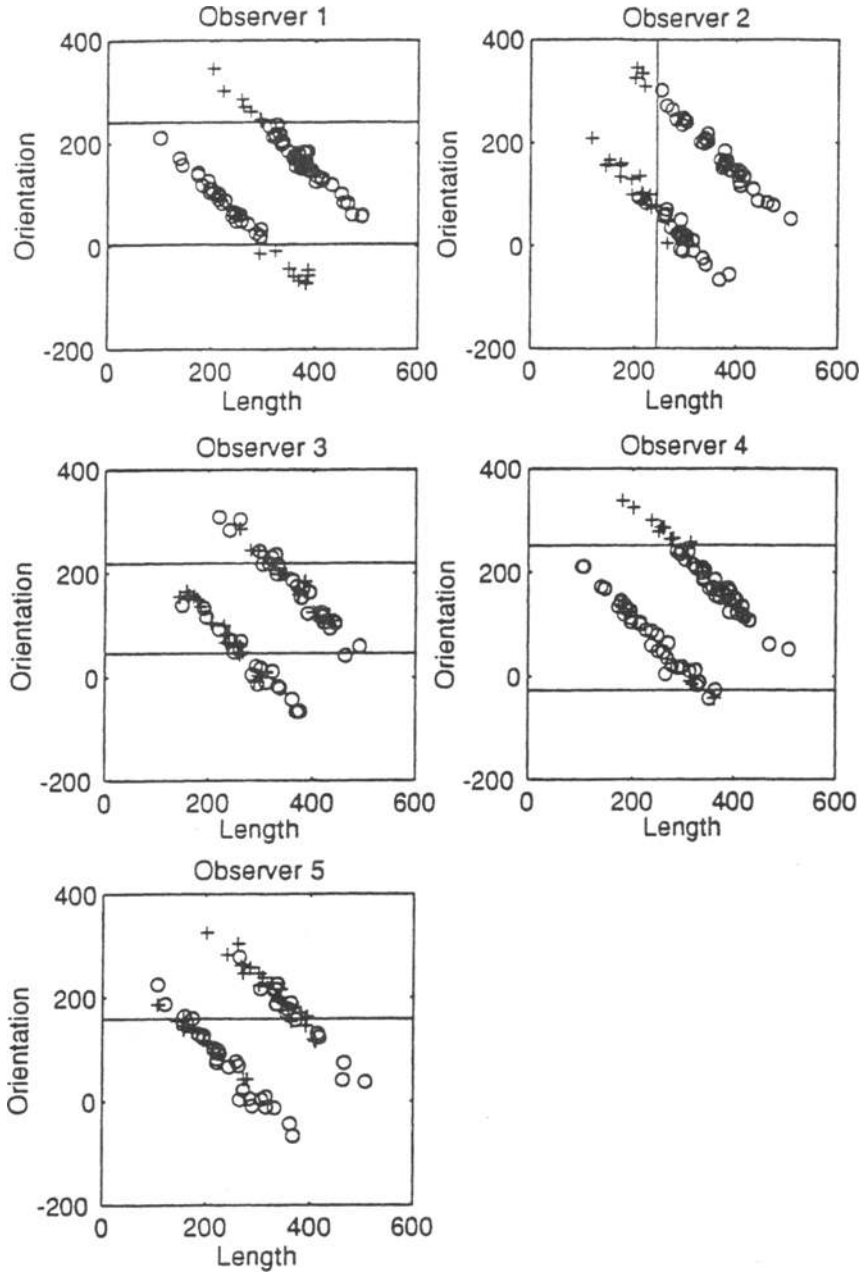

Figure 8. Category responses during the last block for all observers in the diagonal–negative condition of Experiment 1A.

this case, the optimal model uses a unidimensional rule, so Tables 4 and 5 show that, for every response block in the experiment, the data of every observer was best fit by a model that assumed unidimensional responding. The learning that occurred is also apparent. During the last block, the best account of the data is that observers responded optimally. Table 5 indicates, however, that the data collected during 13 of the blocks in the unidimensional conditions was best described by some suboptimal unidimensional rule. In fact, during 5 of these blocks, observers were apparently attending to the wrong dimension.

In the diagonal conditions, the data from the last block were best fit by a model assuming a unidimensional rule for 9 of 10 observers. When all blocks are included, this ratio is 43 of 50. Of the 7 blocks for which the GLC provided the best fit, the best-fitting bound was always closer in slope to a unidimensional rule than to the optimal rule.[6] Two of the 7 best-fitting GLC bounds were almost

Table 4
Number of Times Each Model Was Best Fitting for the Last Block of Data From Experiment 1A

		Unidimensional			
Condition	GLC	Orientation	Length	Interval-Based	Optimal
Unidimensional					
Orientation	0	0	0	0	5
Length	0	0	0	0	5
Diagonal					
Positive	1	3	0	1	0
Negative	0	1	1	3	0

Note—GLC, general linear classifier.

horizontal (differing from horizontal by 0.5° and 10°, respectively), and 4 were nearly vertical (differing from vertical by 6°, 8°, 12°, and 16°). Of the 7, only one differed sharply from a unidimensional bound. This was in the last block for Observer 5 in the diagonal–positive condition (i.e., 23° off vertical; see Figure 7). As can be seen in Figure 7, the best-fitting bound in this case is nearly orthogonal to the optimal bound, so this observer clearly had not discovered the underlying category structure.

There are two interesting possibilities that might explain the significant departure from a unidimensional rule seen in the last block of Observer 5 in the diagonal–positive condition. The first is that this observer might have become aware of the positive correlation between length and orientation that existed for all stimuli in the experiment, even though he or she was never aware of the two separate clusters of stimuli. In this case, a reasonable strategy would have been to use a bound orthogonal to the optimal bound, because this would separate the set of all stimuli into two categories with the property that at least some of the exemplars in each category were far from the decision bound. This possibility is difficult to rule out for this data set, but it fails to account for any of the other data sets that were best described by the GLC.

A second possibility is that this observer tried to use a unidimensional rule on length, but was especially susceptible to the horizontal–vertical illusion. Because of this illusion, vertical lines appear longer than horizontal lines of the same length. In Figure 7, horizontal lines correspond to a value of 0 on the orientation dimension and vertical lines correspond to a value of 250. Suppose the observer sets a criterion on length, with the property that a 45° line that is 300 pixels long is exactly on the bound that separates the two categories. When a horizontal line of length 300 pixels is presented, it will appear shorter than the 45° line, so it will be assigned to the "short" category. On the other hand, when a vertical line of length 300 pixels is presented, it will appear longer than the 45° line, and so it will be assigned to the "long" category. As a result, the bound that best fits such data will have a negative slope. This horizontal–vertical illusion hypothesis is especially attractive since it can also account for the four data sets in which the best-fitting bound was nearly vertical. Specifically, as predicted by the horizontal–vertical illusion, each of these four data sets was best fit by a version of the GLC in which the slope of the decision bound was negative.

Conclusions. These data strongly support the hypothesis that observers in the diagonal conditions never learned the underlying category structure and that, in the absence of noticeable structure, they relied instead on unidimensional or interval-based rules to define the categories. These rule types are similar in that they both require attention to only one dimension. Observers in the unidimensional conditions also used unidimensional rules. However, they eventually learned to select the unidimensional rule that corresponded to the true underlying category structure. Furthermore, there is good evidence that the success of the observers in the unidimensional conditions was not due to the coincidence that they all happened to prefer the optimal rule. If this had been true, accuracy would have been perfect throughout the entire experimental session. Instead, the data indicate that a number of observers first tried a unidimensional rule of the wrong type (e.g., a unidimensional orientation rule in the length condition) and then spontaneously switched to the optimal rule (e.g., see Figure 4 and Tables 4 and 5).

In addition, at the end of their participation, each observer was queried about his or her response strategy. In virtually every case, observers in the unidimensional conditions expressed strong confidence that they had been categorizing the stimuli correctly at the end of the session. In contrast, virtually all observers in the diagonal conditions were confident that they had *failed* to categorize the stimuli correctly. This indicates that observers were able to generate some sort of internal feedback signal, and that they were trying to maximize accuracy. The fact that only the observers in the unidimensional conditions were suc-

Table 5
Number of Times Each Model Was Best Fitting for All Blocks of Data From Experiment 1A

		Unidimensional			
Condition	GLC	Orientation	Length	Interval-Based	Optimal
Unidimensional					
Orientation	0	0	1	6	18
Length	0	4	1	1	19
Diagonal					
Positive	4	14	0	7	0
Negative	3	3	5	14	0

Note—GLC, general linear classifier.

cessful is consistent with the hypothesis that people are *constrained* to use rules that operate on a single stimulus dimension during unsupervised category learning (e.g., unidimensional rules, interval-based unidimensional rules). An interesting question is whether performance would have been better if conditions had been arranged so that the best unidimensional rules had performed even more poorly in the diagonal conditions (i.e., worse than 80% correct). The idea is that this might make it easier for observers to learn that the optimal rule was not unidimensional. However, observers in the diagonal conditions already knew they were performing poorly, yet could do nothing about it, so it seems unlikely that such a manipulation would significantly improve performance.

EXPERIMENT 1B

The observers in Experiment 1A were unable to learn the nondimensional category structures, even with extensive practice. This result contrasts sharply with findings from supervised categorization experiments, which have shown that people can learn some complex (i.e., nonlinear) categorization rules (e.g., Ashby & Maddox, 1992; McKinley & Nosofsky, 1995). Before we conclude that there are some fundamental differences in the constraints on supervised and unsupervised category learning, however, it is important to be convinced that supervised learning is possible in the diagonal conditions of Experiment 1A.

Past research with the randomization technique used in Experiment 1A (i.e., normally distributed categories) suggests that, when given feedback on each trial, people learn to respond almost optimally in conditions in which the optimal bound is linear but not unidimensional, even with highly overlapping categories (Ashby & Maddox, 1990; Maddox & Ashby, 1993). Since no feedback was provided in Experiment 1A, we chose categories that were widely separated in an attempt to make the category structure apparent. Because of the past research with this paradigm, we assumed that with supervision (i.e., with feedback), people would respond optimally in all conditions of Experiment 1A. Certainly, the success of observers in the unidimensional conditions supported this assumption. However, in addition to the use of feedback, there was another important difference between Experiment 1A and the earlier studies (i.e., of Ashby & Maddox, 1990; Maddox & Ashby, 1993). The observers in the earlier studies were trained over multiple days and responded to more than 800 stimuli by the end of training. It was possible, then, that Experiment 1A was a more difficult task than intended, simply because the number of trials was smaller (i.e., 800) and learning did not occur across multiple days. In Experiment 1B, we tested whether the categories used in the diagonal–positive condition of Experiment 1A could be learned with feedback,[7] given the experimental conditions used in Experiment 1A. Thus, Experiment 1B replicated the diagonal–positive condition of Experiment 1A, except that feedback was provided on every trial.

Method

The stimuli used in Experiment 1B were the same as those in the diagonal–positive condition in Experiment 1A. Five observers participated in this experiment. The procedure was the same as that for Experiment 1A, except that the displays were always response terminated, the interstimulus interval was 1.5 sec, observers responded during all 10 blocks, and they received feedback after each of the 800 trials. A short high-pitched tone sounded after each correct response, and a longer low-pitched tone sounded after each incorrect response.

Results and Discussion

Recall that if observers responded optimally, they could achieve 100% accuracy on this task and that during the last block, observers in the corresponding condition in Experiment 1A achieved no better than 76% correct. The percentages of correct responses for the 5 observers during the last block of Experiment 1B were, respectively, 88.8%, 100%, 92.5%, 86.3%, and 98.8%. Thus, the average percent correct during the last block was 93.3%. The average accuracy during the most accurate block was 96%. For each observer, these accuracy values are considerably higher than is predicted for the most accurate unidimensional rule. The actual responses made by each observer during the last block of trials and the best-fitting decision bounds are shown in Figure 9.

The accuracy data suggest that observers responded almost optimally in this task. To test this hypothesis more rigorously, we fit the five models described in the last section to the data from each even-numbered response block (i.e., the same blocks that the models were fit to in Experiment 1A). Of these 25 data sets, the optimal model provided the best fit in 11 cases, the GLC fit best in 9 cases, the interval-based unidimensional model fit best 3 times, and the unidimensional–orientation and unidimensional–length models each fit best once. For every observer, the data from the last block was best fit by either the optimal model (three times) or the GLC (twice). For 4 of the 5 observers, the optimal model provided the best fit to the data from either the 8th or the 10th block of data. Thus, these analyses provide convincing evidence that the category structure used in the diagonal–positive condition in Experiment 1A can be learned when feedback is provided.

EXPERIMENT 2

In Experiment 1A, observers attended to a single stimulus dimension, even when the categories could be separated only by a rule that integrated information across both dimensions. We tried to induce observers to respond optimally by telling them that perfect accuracy was possible. Clearly, these instructions were insufficient. Would any other experimental conditions induce optimal responding in the diagonal conditions of Experiment 1A? Regehr and Brooks (1995) tried a number of manipula-

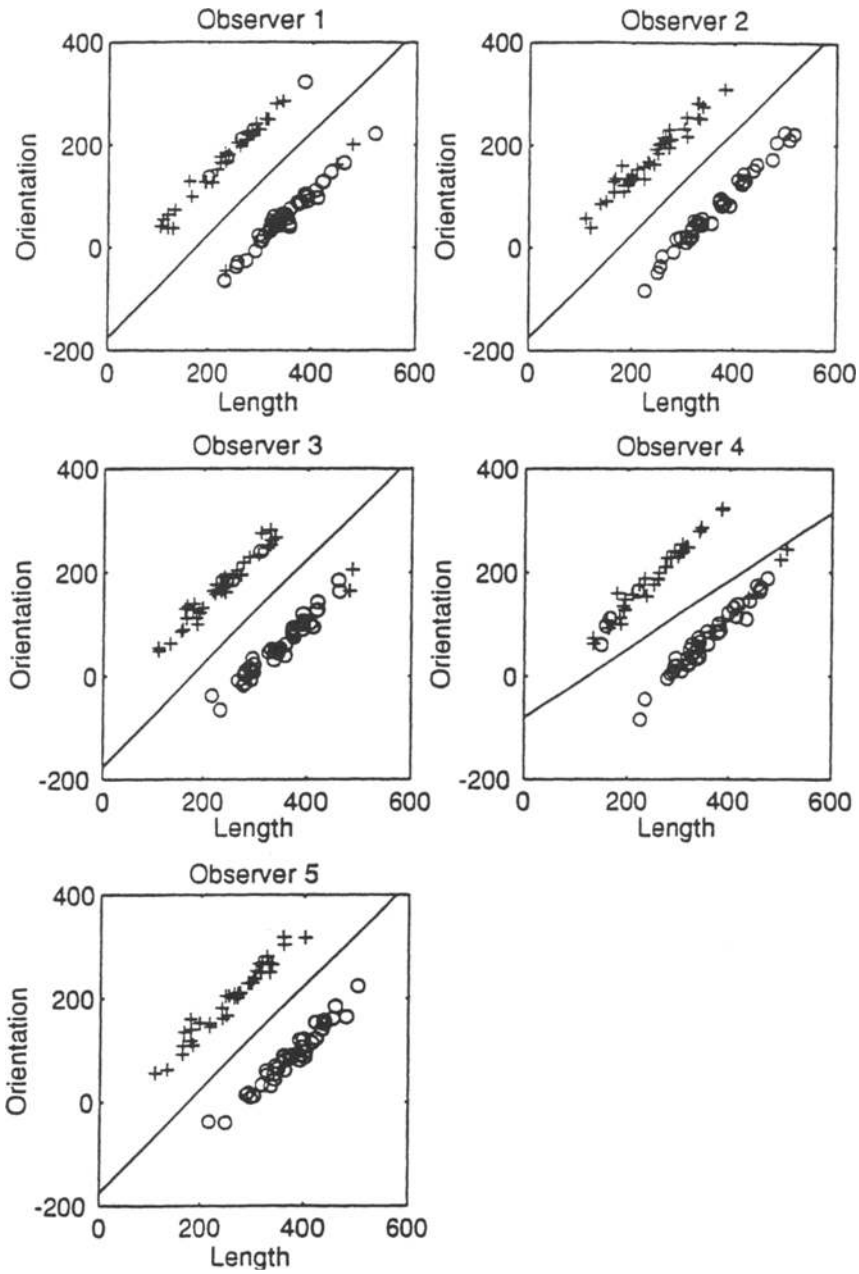

Figure 9. Category responses during the last block for all observers in Experiment 1B.

tions intended to diminish the tendency of observers in unsupervised experiments to rely on a single stimulus dimension. They were finally successful with a procedure in which the two category prototypes were displayed to the observer continuously. Under these conditions, observers appeared to assign stimuli to the category with the most similar prototype. Presumably, the same manipulation would succeed here, because in all four Experiment 1A conditions, the optimal categorization rule was equivalent to a similarity-to-prototype rule. One danger with this intervention, however, was that the task would become one of similarity judgment rather than category learning. Also, in natural settings, category prototypes are virtually never available, so an important practical question to ask was whether we could give general *instructions* to the observers that would induce optimal responding. An obvious possibility was to provide instructions that unidimensional rules are incorrect. If an observer knows in advance that a unidimensional rule is incorrect, he or she need not waste time experimenting with unidimensional rules. If observers are capable of learning and implementing nondimensional rules in the absence of feedback, such instructions should induce optimal responding. Experiment 2 tests this hypothesis.

Table 6
Percent Correct for Each Observer During the Last Response Block of Experiment 2

	Observer					
Condition	1	2	3	4	5	6
Positive	78.8	72.5	73.8	57.5	53.8	—
Negative	85.0	86.3	50.0	76.3	50.0	52.5

Method

Experiment 2 was identical to the diagonal conditions of Experiment 1A in every detail, except for the initial instructions (i.e., separate diagonal–positive and diagonal–negative conditions were run; see Figure 2). In addition, observers in Experiment 2 were also told that the stimuli they were about to see varied in length and orientation *and* that in order to achieve 100% accuracy, they would have to use both the length and orientation information when selecting a response. Five observers participated in the diagonal–positive condition, and 6 participated in the diagonal–negative condition.

Results

The response accuracy of each observer during the last response block is shown in Table 6, and the actual responses and best-fitting decision bounds during the last block are shown in Figures 10 and 11. The results of Experiment 2 paralleled those of Experiment 1A. The average percent correct in the last block was only 67.3% in the diagonal–positive condition and 66.7% in the diagonal–negative condition.

The same five models that had been used in Experiment 1 were fit separately to the data from each observer's five response blocks. In addition, a visual inspection of the data indicated that some observers may have used a conjunctive rule during one or more response blocks (e.g., see Observer 1 in Figure 11). Therefore, in addition to the five models described above, we also fit a series of conjunctive rule models that assumed a decision rule of the following form (Ashby, 1992; Nosofsky et al., 1989):

Respond A if length $< x_1$

AND if orientation $< y_0$, otherwise respond B,

where x_1 and y_0 were free parameters (along with the noise parameter). Four different versions of this model were created by systematically replacing the two "<" signs with all possible combinations of "<" and ">." Models that assume a conjunctive rule are theoretically more similar to models that assume some form of unidimensional rule than to the GLC. This is because, with a conjunctive rule, observers never integrate information from the two stimulus dimensions. Instead they make separate decisions about the two dimensions and then select a response on the basis of the outcomes of these decisions (Ashby & Gott, 1988; Shaw, 1982). In contrast, in the GLC, the stimulus information is integrated (via some linear combination rule) and a response is made on the basis of this integrated value.

The results of fitting these six different model types to the data from each response block for every observer are shown in Table 7. Across the two conditions, the data from only 7 of 55 blocks were best fit by a model which assumed that observers integrated information across the stimulus dimensions (i.e., the GLC). In 6 of these 7 cases, the best-fitting decision bound was closer to a unidimensional rule than to the optimal rule. In 2 of the 7 cases, the best-fitting bound was within 10° of horizontal. In the other 5 cases, the best-fitting bounds were all closer to vertical than to horizontal and in every case had a negative slope. Four of these 5 were within 13° of vertical, and 1 was within 27° of vertical. Thus, one possibility is that during these 5 blocks, the observers were trying to use a unidimensional rule on length but were especially susceptible to the horizontal–vertical illusion. In 48 of the 55 data sets, the best fit was provided by a model that assumed some sort of unidimensional responding (i.e., the unidimensional and interval-based models) or that observers based their responses on separate unidimensional judgments (i.e., the conjunctive model).

Figures 10 and 11 show that during the last response block, the data from 2 observers were best fit by a simple unidimensional rule on orientation, the data from 6 observers were best fit by an interval-based unidimensional rule, and the data from 2 observers were best fit by a conjunctive rule. Only one data set was best fit by a model that assumed integration of information across stimulus dimensions (Observer 2, diagonal–negative condition), and these data are not inconsistent with the hypothesis that this observer tried to use a unidimensional rule but was especially susceptible to the horizontal–vertical illusion. Thus, the modeling analysis strongly supports the conclusion that none of the observers learned the category structure and that instead of integrating information, they tended to focus on a single stimulus dimension.

Experiment 2 indicates that, even when observers are explicitly encouraged to use information from both dimensions, they fail to integrate information across separable stimulus dimensions in the absence of feedback. Encouraging observers to use information from both dimensions did draw some observers away from the use of simple unidimensional rules. However, rather than re-

Table 7
Number of Times Each Model Was Best Fitting for All Blocks of Data From Experiment 2

			Unidimensional			
Condition	Optimal	GLC	Orientation	Length	Interval Based	Conjunctive
Diagonal positive	0	3	5	1	15	1
Diagonal negative	0	4	4	5	13	4

Note—GLC, general linear classifier.

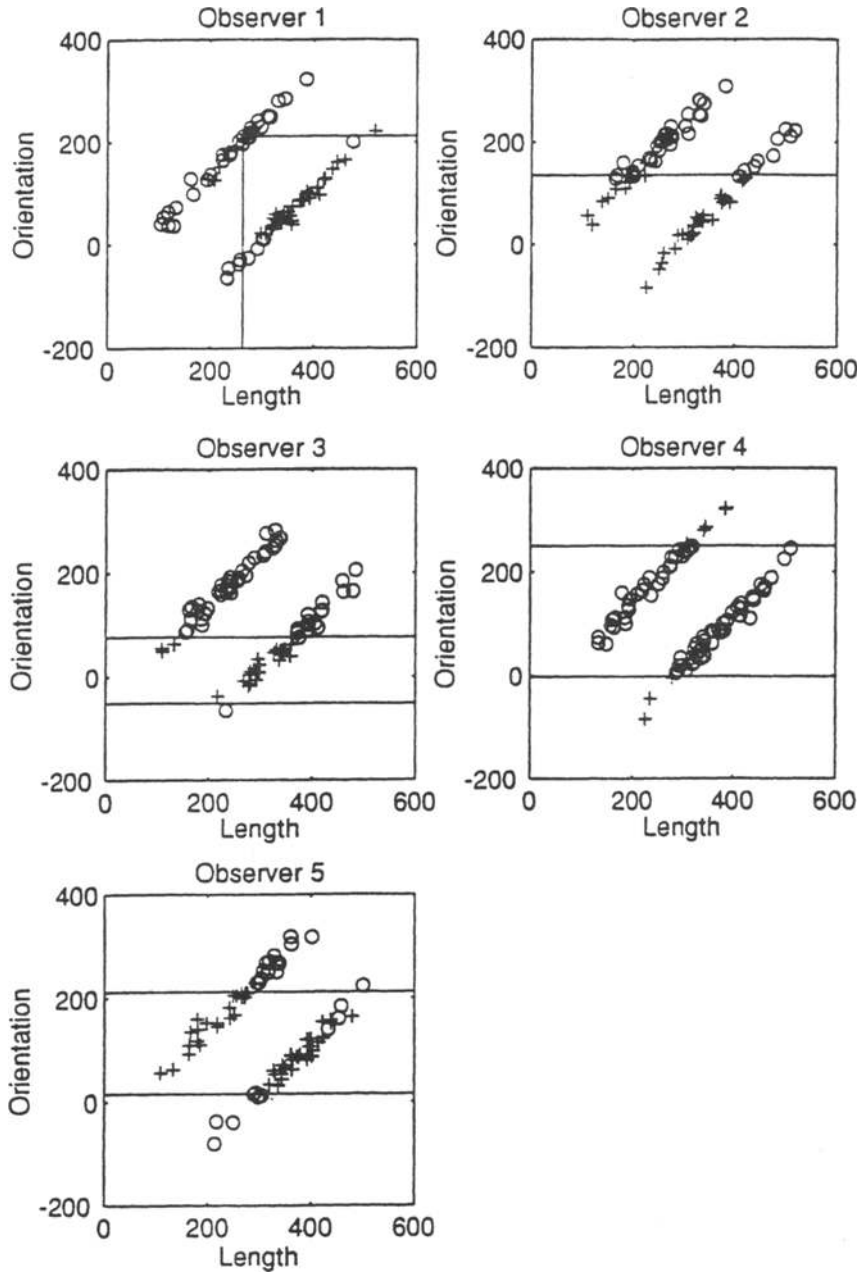

Figure 10. Category responses during the last block for all observers in the diagonal–positive condition of Experiment 2.

sponding optimally, these observers used conjunctive rules—that is, they used separate unidimensional rules to classify each stimulus component before selecting a response.

GENERAL DISCUSSION

The observers in Experiments 1 and 2 learned to respond optimally without feedback when they were given enough practice and were shown categories that were defined by coherent clusters of stimuli, but only when the optimal rule was unidimensional. When the optimal rule was diagonal, they responded with unidimensional or conjunctive rules, even when told that both stimulus dimensions must be used, and even under conditions in which optimal performance was quickly achieved when feedback was provided. Of course, our results do not rule out the possibility that, with enough experience, observers would eventually have discovered the correct category structure in the diagonal conditions. Nor do they rule out the possibility that observers in the diagonal conditions would have been more successful with some qualita-

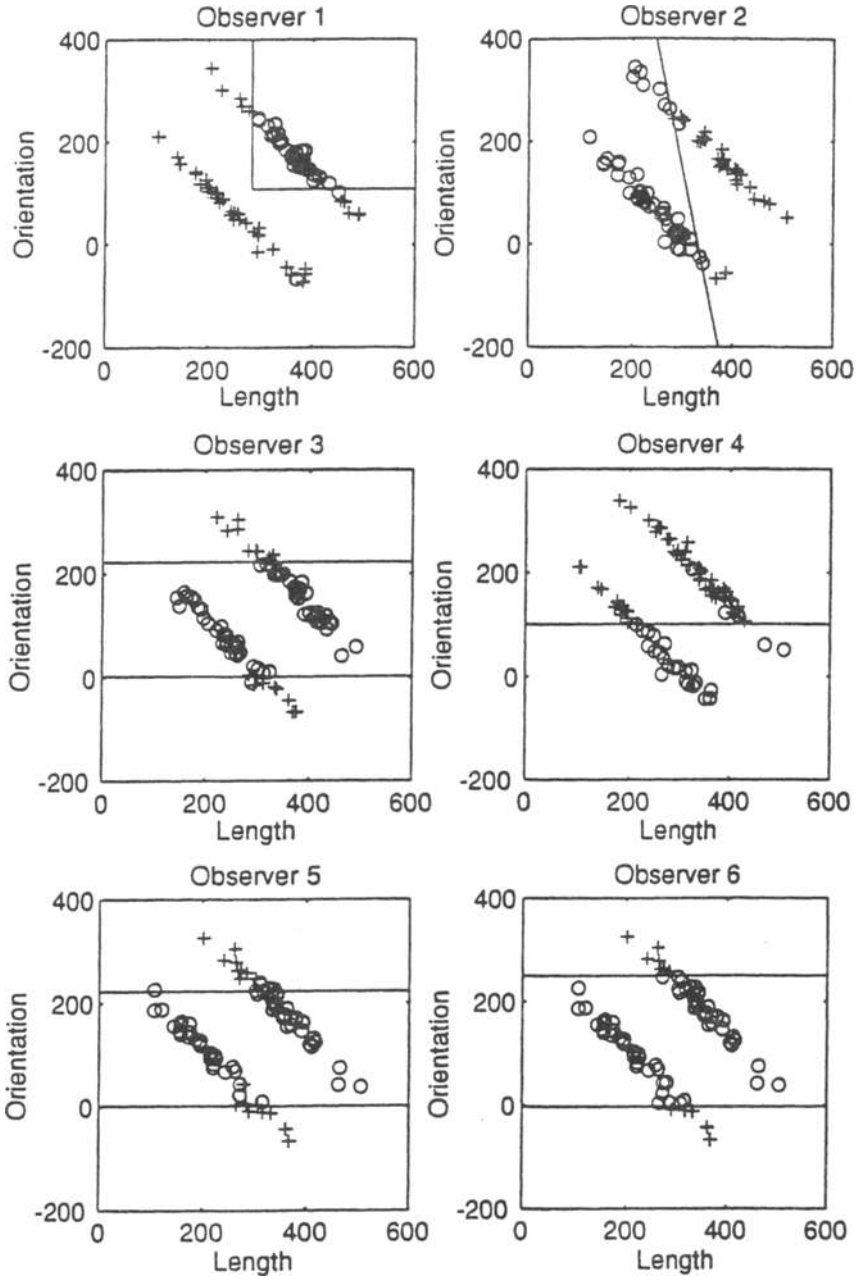

Figure 11. Category responses during the last block for all observers in the diagonal–negative condition of Experiment 2.

tively different stimuli. Even so, our results demonstrate a striking bias in favor of rules that operate on one stimulus dimension at a time. In a number of supervised categorization studies, it has been reported that people are often biased toward unidimensional rules (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; McKinley & Nosofsky, 1996), so there is considerable converging evidence that unidimensional rules have some privileged status in human categorization behavior.

To perform optimally in the unidimensional conditions, observers had to determine which of the two stimulus dimensions was critical, and they had to learn the correct value of the response criterion (i.e., the intercept of the vertical or horizontal decision bound). Unfortunately, our results shed little light on how either of these problems was solved. For example, we cannot rule out the possibility that observers were drawn to the dimension with less overall variability or to a criterion setting that was at the midpoint of the stimulus values on the attended dimension. On the other hand, the randomization technique used in the studies reported above makes it easy to answer these questions. For example, within-category variance

on the critical stimulus dimension could be increased (in the unidimensional conditions) to make the overall variability larger on the critical dimension than on the irrelevant dimension. If people are drawn to the stimulus dimension with less overall variability, then learning should fail in this condition.

In comparison with the number of models of supervised categorization, there are relatively few models of unsupervised categorization. For example, although exemplar theory has been widely used to model categorization data from supervised experiments (see, e.g., Brooks, 1978; Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986), there is no well-developed exemplar-based account of unsupervised categorization (e.g., Billman, 1992; Wattenmaker, 1992). Perhaps the two best-known models of unsupervised categorization are Ahn and Medin's (1992) two-stage model and Anderson's (1991) rational model (for another interesting model,[8] see Billman & Heit, 1988).

The two-stage model of Ahn and Medin (1992) assumes that during unsupervised categorization, observers apply a unidimensional rule to the most salient stimulus dimension. If this strategy fails to separate the stimuli into coherent clusters, the stimuli with intermediate values on the critical dimension are compared with the stimuli with extreme values. The intermediate stimuli may then be reassigned to categories on the basis of these similarity computations. By placing special emphasis on unidimensional rules, the two-stage model seems to be in a good position to account for the major results of Experiment 1A. Unfortunately, however, the model was never formalized, so it is unclear what predictions it would make for the diagonal conditions. Also, it is strictly a model of unsupervised categorization, so it is not clear how the two-stage model would account for the differences between the data from Experiment 1B and the data from the diagonal–positive condition in Experiment 1A.

In Anderson's (1991) rational model, the observer partitions the stimuli into clusters in a sequential fashion. To create the partitions, on each trial the observer computes the probability that the presented stimulus is a member of each existing cluster, together with the probability that the stimulus is a member of a new cluster. The stimulus is assigned either to an existing cluster or to a new cluster, depending on which of these probabilities is highest. This algorithm is order sensitive, since a different set of clusters could be formed if the order of stimulus presentation was changed. In the rational model, the clusters that are formed are required to have values on different stimulus dimensions that are statistically independent. Except for this constraint, the model assumes that observers are approximately optimal in their use of available information during unsupervised categorization. Alternatively, it is essentially a multiple prototype model, with each prototype defining its own cluster of stimuli.

What predictions does the rational model make for the experiments reported in this article? In the unidimensional conditions, statistical independence holds within each category, so presumably the rational model correctly predicts optimal responding in these conditions. In the diagonal conditions, however, independence fails. How the model responds to this violation of independence depends on the value of its coupling parameter. If the coupling parameter is large, the observer will perceive only a single cluster of stimuli in the experiment and so will respond randomly. The problem with this explanation is that such a large value of the coupling parameter would cause the observer also to respond randomly in the unidimensional conditions. With a smaller value of the coupling parameter, the rational model predicts that the observer would perceive the stimuli in the diagonal conditions as belonging to a number of clusters, with the property that within each cluster, statistical independence holds. Such clusters would have a circular shape, or they would be ellipses with either a horizontal or a vertical orientation. The model could predict a difference between the unidimensional and diagonal conditions only if at least some of the clusters that emerged in the diagonal conditions were ellipses that included members of the two separate categories. This might occur if the value of the coupling parameter was large enough to ignore the vertical or horizontal distance between categories (see Figure 2). However, the vertical distance between categories in the unidimensional–orientation condition and the horizontal distance between categories in the unidimensional–length condition are smaller than the analogous vertical and horizontal distances in the diagonal conditions. Thus, it seems that the rational model could not account for the qualitative difference in performance observed in the diagonal and unidimensional conditions with the same value of the coupling parameter.

Given that our results present problems for current theories of unsupervised category learning, it makes sense to ask what basic assumption should be used to construct such a theory. One obvious possibility is that unsupervised learning can occur in simple, but not complex, tasks. The idea here is that the unidimensional conditions are considerably simpler than the diagonal conditions, and this complexity difference is the critical factor, rather than the unidimensional versus nondimensional distinction, per se. The argument that the unidimensional conditions are simpler than the diagonal conditions seems reasonable, since the unidimensional structures require observers to attend to only one dimension, whereas the diagonal structures require observers to attend and integrate information from two dimensions. Also, when feedback is given, observers learn more quickly in the unidimensional conditions than in the diagonal conditions (Ashby et al., 1998). One barrier to testing the simplicity hypothesis rigorously, however, is that, in contrast to the concept of a "unidimensional rule," there is no generally accepted definition of a "simple rule." Developing such a definition and testing whether simplicity predicts unsupervised learning should be a high priority for future research.

Another possibility is suggested by a recent neuropsychological theory of category learning, called COVIS

(competition between verbal and implicit systems), which assumes that category learning is a competition between separate explicit and implicit categorization systems (Ashby et al., 1998). The explicit system is a logical reasoning system under conscious control that engages in a systematic process of hypothesis testing (as postulated, e.g., by Bruner, Goodnow, & Austin, 1956) or theory construction and testing (as postulated, e.g., by Murphy & Medin, 1985). Explicit rules were defined operationally as those rules that are easy to describe verbally (hence the acronym COVIS). The implicit system is assumed to engage in a form of procedural learning. COVIS assumes that the explicit system initially dominates, presumably because it is controlled by consciousness. With feedback and experience, however, the potential of the implicit system for superior performance often eventually overcomes the initial bias in favor of the explicit system. From the perspective of COVIS, the main difference between the unidimensional and diagonal conditions of Experiment 1A is that the optimal rules in the unidimensional conditions were explicit (e.g., they could easily be verbalized), whereas the optimal rules in the diagonal conditions were not. For example, in the unidimensional–length condition, the optimal rule could easily be verbalized as "give one response if the length exceeds some criterion value, and give the other response if it does not." As mentioned above, at the end of their participation, every observer was queried about his or his response strategy. Every observer in all four conditions of Experiment 1A correctly described the unidimensional rule that best fit his/her data. In the diagonal conditions, however, the optimal rule had no salient verbal description. For example, a verbal description of the optimal rule in the diagonal–positive condition might be "give one response if the orientation exceeds the length, and give the other response if it does not." However, orientation and length are in different units, so it is not clear what such a rule would mean. None of the observers in Experiment 1B described their behavior in such language, even though their data were well described by such a rule. Thus, COVIS predicts that in Experiment 1B, the observer's explicit system will try to discover the most accurate unidimensional rule (since these are the only rules that it can easily describe[9]) at the same time as the implicit system is performing an unconstrained search for the optimal rule among all possible (linear) bounds.

An unsupervised categorization version of COVIS has not been developed, but given its emphasis on explicit rules, the present results indicate that COVIS might provide a promising foundation on which to build a powerful model of unsupervised categorization.[10] In the past, the literature on unsupervised categorization has emphasized unidimensional rules. COVIS suggests that this is because unidimensional rules are almost always the most salient explicit rules. However, many other rules are also explicit, so COVIS predicts that occasionally, people should adopt rules that are not unidimensional. In fact, we found evidence of this in Experiments 1A and 2. In many response blocks (see Tables 5 and 7), observers in these experiments used an interval-based unidimensional rule of the following type: "Respond A if the orientation is between horizontal (i.e., 0°) and vertical (i.e., 180°), otherwise respond B." Although this rule involves only one dimension, it seems more complex than the simple unidimensional rules that have been identified in the literature. Even more striking is the fact that the data from five of the response blocks in Experiment 2 were best fit by a conjunctive model which assumes that observers used a rule of the following type: "Respond A if length $< x_1$ AND if orientation $< y_0$, otherwise respond B." This rule is explicit, but it is very different from the unidimensional rules identified in the unsupervised category learning literature. The present results (and COVIS) are consistent with the hypothesis that the success of unsupervised category learning depends on whether the optimal rule is explicit, rather than on whether it is unidimensional. Obviously, however, much more work is needed to test this idea. Most important would be experiments in which the optimal rule is explicit, but not unidimensional.

CONCLUSIONS

In the studies described above, category learning was strikingly different, depending on whether feedback was provided. With feedback, observers in Experiment 1B learned a diagonal rule that required integrating information from perceptually separable stimulus dimensions. This result is consistent with a number of similar findings in the literature (e.g., Ashby & Maddox, 1990, 1992; McKinley & Nosofsky, 1995). Without feedback, observers in Experiments 1A and 2 were only able to learn unidimensional rules, and they persisted in using unidimensional and conjunctive rules even when such rules failed to separate the stimuli into obvious clusters.

Although these results provide new insights into the limits on unsupervised category learning, they also raise many interesting new empirical questions. Important among these are the following. Do the constraints on category learning observed in our studies generalize to other, qualitatively different stimuli (e.g., those constructed from integral dimensions)? What is the nature of unsupervised criterion learning? What role do simplicity and verbalization play in unsupervised category learning? We believe that the randomization technique used here provides an excellent vehicle from which to attack these important questions.

Finally, there are many practical implications of this work. Most of the hundreds of categorization responses that we make every day are unsupervised. Children (and adults) learning about categories often receive either no feedback or feedback that is inaccurate or untrustworthy (e.g., from siblings and playmates). Our results suggest that in such cases, learning will still often occur. However, the rules that are learned will tend to be unidimensional (or perhaps conjunctive). With widely separated categories, such rules might work perfectly. In many cases, however,

the rules that are learned will correctly assign most, but not all, exemplars to their proper category. In such cases, the person may appear to understand the differences among a particular set of contrasting categories, but on some critical subset of examples, he or she will fail consistently.

REFERENCES

AHN, W.-K., & MEDIN, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, **16**, 81-121.

AIKEN, L. S., & BROWN, D. R. (1971). A feature utilization analysis of the perception of pattern class structure. *Perception & Psychophysics*, **9**, 279-283.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

ALFONSO-REESE, L. A. (1996). *Dynamics of category learning*. Unpublished doctoral dissertation, University of California, Santa Barbara.

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

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

ASHBY, F. G., ALFONSO-REESE, L. A., TURKEN, A. U., & WALDRON, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, **105**, 442-481.

ASHBY, F. G., & GOTT, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.

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

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

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

ASHBY, F. G., & MADDOX, W. T. (1998). Stimulus categorization. In M. H. Birnbaum (Ed.), *Handbook of perception & cognition: Judgment, decision making, and measurement* (pp. 251-301). San Diego: Academic Press.

BILLMAN, D. (1992). Modeling category learning and category use: Representation of processing. In B. Burns (Ed.), *Percepts, concepts, and categories: The representation and processing of information* (pp. 414-448). New York: Elsevier.

BILLMAN, D., & HEIT, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, **12**, 587-625.

BILLMAN, D., & KNUTSON, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Human Learning & Memory*, **22**, 458-475.

BOSTER, J., & D'ANDRADE, R. (1989). Natural and human sources of cross-cultural agreement in ornithological classification. *American Anthropologist*, **91**, 132-142.

BROOKS, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

BRUNER, J. S., GOODNOW, J., & AUSTIN, G. (1956). *A study of thinking*. New York: Wiley.

CLAPPER, J. P., & BOWER, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 27, pp. 65-108). San Diego: Academic Press.

CLAPPER, J. P., & BOWER, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 443-460.

ESTES, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.

EVANS, S. H., & ARNOULT, M. D. (1967). Schematic concept formation: Demonstration in a free sorting task. *Psychonomic Science*, **9**, 221-222.

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.

FUKUNAGA, K. (1990). *Statistical pattern recognition* (2nd ed.). New York: Academic Press.

HOMA, D., & CULTICE, J. C. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 83-94.

IMAI, S., & GARNER, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, **69**, 596-608.

JULESZ, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, **290**, 91-97.

JULESZ, B., & BERGEN, J. R. (1983). Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal*, **62** (Pt. 2), 1619-1645.

MADDOX, W. T., & ASHBY, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, **53**, 49-70.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1995). Investigations of exemplar and decision bound models in large-size, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 128-148.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 294-317.

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

MEDIN, D. L., WATTENMAKER, W. D., & HAMPSON, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, **19**, 242-279.

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

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

NOSOFSKY, R. M., CLARK, S. E., & SHIN, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 282-304.

NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.

REGEHR, G., & BROOKS, L. R. (1995). Category of organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 347-363.

ROSS, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1249-1265.

SHAW, M. L. (1982). Attending to multiple sources of information: I. The integration of information in decision-making. *Cognitive Psychology*, **14**, 353-409.

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

TAKANE, Y., & SHIBAYAMA, T. (1992). Structures in stimulus identification data. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 335-362). Hillsdale, NJ: Erlbaum.

TVERSKY, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, **79**, 281-299.

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

WATTENMAKER, W. D. (1992). Relational properties and memory-based category construction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1125-1138.

WICKENS, J. (1993). *A theory of the striatum*. New York: Pergamon.

NOTES

1. By "coherent cluster," we mean that such clusters are readily apparent under visual inspection. For example, it is obvious that there are two coherent clusters of stimuli in each experimental condition shown in Figure 2. A more rigorous definition could be derived from the object recognition literature, where it has been argued that low spatial frequency channels play an important role in object segmentation (e.g., Julesz, 1981; Julesz & Bergen, 1983). As the high spatial frequencies are gradually removed from Figure 2, a point is eventually reached at which only two objects (i.e., category "blobs") remain in each experimental condition. Thus, the Figure 2 conditions each contain two coherent clusters. In contrast, in Figure 1, this same process initially yields eight blobs (i.e., one for each stimulus), which at some point simultaneously merge into a single blob. Thus, depending on one's criterion, Figure 1 contains either one or eight coherent clusters of stimuli, but it does not contain two clusters.

2. Perhaps the studies that came closest to these criteria were reported by Fried and Holyoak (1984), Homa and Cultice (1984), and McKinley and Nosofsky (1995). We will discuss the former two studies in the next section. The McKinley and Nosofsky (1995) study was a sidebar to a more general investigation of the ability of people to learn categories of stimuli constructed from two continuous-valued dimensions under supervised conditions. Each category was composed of separate clusters, or subcategories. On the last day of one experiment, observers were told that each category was composed of two subcategories, and they were asked to make an unsupervised subcategorization response after making their supervised categorization response. Although the unsupervised categorization data were not presented, McKinley and Nosofsky (1995) reported that observers used unidimensional rules, even though the optimal subcategorization rule was a quadratic function of the dimensional values.

3. With the possible exception that there might be a natural tendency to equate the size (i.e., the cardinality) of the contrasting categories. Even so, this requirement places no constraint on the *form* of the decision rule.

4. For example, a sorting based on overall similarity to the two high-frequency stimuli produces a different category structure from one in which the categories are constructed under the constraint that all members of the same category share a common value on one of the stimulus dimensions.

5. It is important to note that from the observer's perspective, a line of 0° orientation is identical to a line of 180° orientation. Thus, the psychological representation of length and orientation lies on a cylinder (with the orientation dimension wrapping back around on itself). To minimize this problem, we restricted the range of orientation to be less than 180° (500 orientation units in Figure 2).

6. More specifically, the amount of rotation needed to align the best-fitting bound with the nearest unidimensional bound was always less than the amount of rotation needed to align it with the optimal bound.

7. Only the diagonal–positive condition was run, because Maddox and Ashby (1993) found this condition to be more difficult for observers to learn than the diagonal–negative condition.

8. Billman and Heit (1988) developed their model for application to tasks in which the category exemplars vary on many discrete-valued dimensions. It is not immediately clear how to generate predictions from the model for the experiments reported here (since the category exemplars in our experiments varied on two continuous-valued dimensions).

9. Of course, many other rule types can be described verbally, including any rules that apply logical operations to the two separate dimensions (e.g., conjunctive rules, disjunctive rules). However, Alfonso-Reese (1996) found that, when no special instructions are given, rules of this more complicated type have extremely low salience. As a result, when one is deriving predictions from COVIS for Experiment 1, conjunctive and disjunctive rules can be safely ignored. It seems likely, however, that the Experiment 2 instructions to use both stimulus dimensions could increase the salience of conjunctive rules. Therefore, in applications of COVIS to Experiment 2, it is likely that a more complex model of the verbal system will be needed.

10. COVIS assumes that the striatum (a structure in the basal ganglia) is a key structural component of the implicit system. The best available evidence suggests that learning in the striatum is mediated by a dopamine-based reward signal (e.g., Wickens, 1993). Thus, in the absence of feedback, it is expected that striatal learning will be severely impaired.

(Manuscript received February 13, 1998;
revision accepted for publication August 21, 1998.)