Typicality in logically defined categories: Exemplar-similarity versus rule instantiation

ROBERT M. NOSOFSKY

Indiana University, Bloomington, Indiana

A rule-instantiation model and a similarity-to-exemplars model were contrasted in terms of their predictions of typicality judgments and speeded classifications for members of logically defined categories. In Experiment 1, subjects learned a unidimensional rule based on the size of objects. It was assumed that items that maximally instantiated the rule were those farthest from the category boundary that separated small and large stimuli. In Experiment 2, subjects learned a disjunctive rule of the form " x or y or both." It was assumed that items that maximally instantiated the rule were those with both positive values (x and y). In both experiments, the frequency with which different exemplars were presented during classification learning was manipulated across conditions. These frequency manipulations exerted a major impact on subjects' postacquisition goodness-of-example judgments, and they also influenced reaction times in a speeded classification task. The results could not be predicted solely on the basis of the degree to which the rules were instantiated. The goodness judgments were predicted fairly well by a mixed exemplar model involving both relative-similarity and absolute-similarity components. It was concluded that even for logically defined concepts, stored exemplars may form a major component of the category representation.

What is the nature of the mental representation that results when people learn categories through induction from individual examples? Cognitive psychologists have suggested different answers to this question, depending on the kinds of categories that are learned. In the case of logically defined categories structured according to simple rules, it is generally assumed that people formulate and test hypotheses concerning the rules that determine category membership. Categorization decisions are based on whether or not an item satisfies the hypothesized rule. By contrast, in the case of ill-defined categories, the category representation may consist of a prototype, feature frequency counts, or memories of individual exemplars.

A natural question is whether classification involving logically defined and ill-defined categories can be explained within a single framework. Nosofsky (1984, 1986, 1987, 1989; Nosofsky, Clark, & Shin, 1989) proposed and tested a quantitative model that has been successful in the domains of both logically defined and ill-defined categorization. The model, which is a generalized version of the context model of classification proposed by

Medin and Schaffer (1978), is based on the idea that classification decisions are governed by similarity comparisons with individually stored category exemplars.

A critical component assumption of the generalized context model (GCM) that has enabled it to predict classification for logically defined categories is that similarities among exemplars are modifiable by selective attention. Selective attention is represented in the GCM in terms of the stretching and shrinking of distances in a psychological space (Carroll & Wish, 1974). The idea is illustrated in Figure 1A, which shows a set of 16 stimuli varying along two continuous dimensions (angle and size). Locations labeled with a 1 represent training exemplars assigned to Category 1, and likewise for Category 2. A salient rule for describing the categories is that stimuli with low angles are assigned to Category 1, and stimuli with high angles to Category 2. According to this rule, transfer stimulus T should tend to be classified in Category 1. An "overall-similarity" exemplar model predicts the opposite, because transfer stimulus T is highly similar to a Category 2 training exemplar. However, if people attended selectively to the relevant angle dimension, the psychological space would be "stretched" along the angle dimension and "shrunk" along the size dimension (Figure 1B). Now T becomes more similar to the training exemplars of Category 1, allowing the exemplar model to predict the rule-described behavior.

Although the exemplar model has yielded accurate quantitative predictions of classification performance for a wide variety of rule-described structures (Nosofsky, 1984, 1986, 1987, 1989; Nosofsky et al., 1989), a natural reaction to the illustration in Figure 1 is that the exem-

This work was supported by Grant BNS 87-19938 from the National Science Foundation to Indiana University. I thank Hyun Jung Shin for his assistance in programming Experiment 2 and for conducting some of the statistical analyses. The work also benefited from discussions with Greg Ashby and Doug Medin, who provided useful criticism. Finally, for their extensive critiques of earlier submitted versions of the manuscript, I thank Jim Chumbley and several anonymous reviewers. Correspondence should be addressed to Robert Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405.

Figure 1. Schematic rule-described category structure for illustrating an application of the selective-attention exemplar model. (From "Rules and exemplars in categorization identification, and recognition" by R. M. Nosofsky, S. E. Clark, and H. J. Shin, 1989, *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 15, 282-304. Copyright 1989 by the American Psychological Association. Reprinted by permission.)

plar model is mimicking the simple rule. In other words, one's intuition may be that people really represent the category in terms of the rule that "low angles signal Category 1 and high angles signal Category 2," with the exemplar-similarity account being viewed as post hoc. The thesis that is advanced and tested experimentally in the present article, however, is that aspects of the category representation may be *better* characterized in terms of the selective-attention exemplar model than in terms solely of the simple rule.

The design of Experiment 1 is illustrated in Figure 2. The stimuli are circles varying along two continuous dimensions: size of circle and angle of radial line. Stimuli

enclosed by triangles are members of Category A, whereas stimuli enclosed by squares are members of Category B. The obvious rule is that small stimuli (value less than or equal to 2 on the size dimension) belong in Category A and large stimuli in Category B.

The main experimental manipulation is that during training, presentation frequencies of individual exemplars are varied across conditions. Researchers have shown that in learning ill-defined categories, people are sensitive to differential frequencies of individual exemplars (Barsalou, 1981; Estes, 1986b; Nosofsky, 1988b, in press). The aim in the present experiment is to use this finding to advantage to contrast the predictions of the exemplar model and

Figure 2. Design for Experiment 1. Stimuli enclosed by triangles are members of Category A, whereas stimuli enclosed by squares are members of Category B. In some conditions, Stimuli 15 and 16 are not presented during training. Stimuli 13 and 14 are never presented during training. The figure illustrates the psychological spacings between dimension values as determined in a similarity-scaling study in which subjects identified the stimuli. The actual physical values of Angles 1–5 were 5°, 30°, 45°, 60°, and 85°, respectively; and the physical values of Sizes 1–5 (length of radial line) were 5.50, 6.25, 7.25, 8.00, and 19.00 mm, respectively.

a simple rule-based model for the case of the logically defined category shown in Figure 2.

In two of the main conditions, only Exemplars 1–12 are presented during training, whereas the very large stimuli, Exemplars 15 and 16, are never presented. Following training, postacquisition “goodness-of-example” (or “typicality”) ratings are collected to probe the nature of people’s category representations. The central argument advanced here is the following: To the extent that people’s representation of the categories is solely the rule that “small stimuli belong in Category A and large stimuli in Category B,” it is expected that the larger the stimulus, the better an example of Category B it should be judged. That is, goodness-of-example ratings should be determined by *the degree to which the rule is instantiated*—in the present case, by the distance of an item from a category boundary that partitions small and large stimuli. Thus, although Transfer Stimuli 15 and 16 are never presented during training, they best instantiate the rule and should be rated as the best examples of the category. By contrast, to the extent that people’s category representations consist of individually stored exemplars, it is expected that Stimuli 15 and 16 will have diminished goodness ratings.

Motivation for the Rule Model

It should be emphasized from the outset that the present experiment is designed to contrast the predictions of the exemplar model with those of *one particular* logically defined rule-based model. It would be absurd to claim that the exemplar model is being contrasted with the entire class of rule-based models, because in its most general sense a “rule” can be essentially anything. The purpose in this section is to review previous theoretical and empirical work suggesting that the particular rule-based model under consideration is a salient, well-motivated, and serious candidate.

At the heart of the present rule model are two fundamental assumptions. The first is that a single category boundary is established to partition the members of Categories A and B. This boundary corresponds to setting a criterion on the dimension of size such that objects whose size exceeds the criterion are classified in Category B, with the remaining objects classified in Category A. The second assumption is that people’s goodness-of-example judgments (with respect to a given category) will be a monotonically increasing function of the distance of an object from the boundary.

The assumption that observers establish decision criteria to partition perceptual spaces into response regions is ubiquitous in experimental psychology. It is a core assumption of signal detection theory and of Thurstonian approaches to modeling classification (Green & Swets, 1966; Thurstone, 1927). Although most applications of signal detection theory have been in unidimensional domains, there has been extensive recent work in generalizing the theory for application in multidimensional domains (e.g., Ashby & Gott, 1988; Ashby & Townsend, 1986).

The idea that an observer establishes decision criteria in multidimensional spaces is an extremely general one, and additional assumptions are necessary if it is to serve as a useful predictive theory of classification behavior. At the heart of the present rule model is the assumption that the choice of decision criterion is strongly influenced by a “simplicity metric.” If an extremely simple (and psychologically salient) rule is available for partitioning the members of contrasting categories, it will be favored over more complex rules.

One way to formalize the construct of simplicity of rules is to do so in terms of versions of set-theoretic logic (e.g., Neisser & Weene, 1962; Nosofsky et al., 1989; Shepard, Hovland, & Jenkins, 1961; Suppes, 1957). In the present case, illustrated in Figure 2, we assume that the basic elements in the system are the levels on the component dimensions. (We denote levels on Dimension 1 with a single prime, and levels on Dimension 2 with a double prime.) In addition to the standard logical operators “and” ($\wedge$), “or” ($\vee$), and “not” ($\neg$), for the present continuous-dimension stimuli we use $>$, $\geq$, $<$, and $\leq$. Then membership in Category B can be described by the simple set-theoretic string, $B: \geq 3''$; that is, members of

Category B are those with values greater than or equal to 3 on the size dimension. Alternative descriptions of Category B are also available—for example, $B: 3'' \vee 4''$, $B: \geq 3'' \wedge < 5''$, $B: \geq 3'' \wedge (2' \vee 3' \vee 4')$, and so forth. But if the simplicity metric (length of set-theoretic string) underlies the choice of decision criterion, then the rule $B: \geq 3''$ is predicted to operate.

The assumption of some form of simplicity metric has been made in a vast amount of previous psychological research concerned with the learning and representation of logically defined concepts. Examples of explicit statements of the idea come from Neisser and Weene (1962) and Shepard et al. (1961), who tested and found support for the hypothesis that the difficulty of learning different concepts was directly related to the length of the shortest set-theoretic string required to describe the concepts. In closely related work, Leeuwenberg (1971) and Restle (1979) found support for the hypothesis that people's perceptions of ambiguous patterns were systematically biased toward those patterns that enabled the most economical abstract description.

In a recent extension of signal-detection theory to multi-dimensional domains, Ashby and Gott (1988) also posited the role of a simplicity metric. Across three experiments, subjects learned to classify two-dimensional perceptual stimuli into two categories. Ashby and Gott observed patterns of performance consistent with the idea that subjects had established a simple linear boundary in the two-dimensional space, such that objects falling in "Region A" of the space were always classified in Category A. Indeed, on the basis of their data, Ashby and Gott concluded:

A plausible model suggested by the experiments reported here is that subjects first choose very simple deterministic decision rules that are easy to implement and place a minimal burden on available cognitive capacities. . . . If, by whatever the relevant criterion, the subject decides that this rule produces satisfactory performance, then it will be retained even if nonoptimal. On the other hand, if the simple rule leads to inadequate performance, the subject will expend a little more effort and adopt a more accurate rule. (pp. 51-52)

In the case of the present category structure, illustrated in Figure 2, the logical rule $B: \geq 3''$ is simple, and it appears to be psychologically salient and easy to implement. With regard to the discrimination of the classes of objects, it enables not only adequate performance, but essentially optimal performance. Given the previous literature on concept learning and signal detection which points to the role of simplicity in the formation of rules, it is clearly a well-motivated rule candidate.

The second main assumption embodied in the present rule model is that subjects' goodness-of-example judgments will be a monotonically increasing function of an object's distance from the rule boundary. Again, similar hypotheses based on distance-from-boundary are ubiqui-

tous in the psychological literature. For example, numerous investigators have assumed that classification confidence ratings are an increasing function of the distance of an object from a category boundary, or that classification reaction times are a decreasing function of distance of an object from the boundary (e.g., Cartwright, 1941; Emmerich, Gray, Watson, & Tanis, 1972; Festinger, 1943; Green & Swets, 1966; Norman & Wickelgren, 1969; Pike, 1973; Smith, 1968; Thomas, 1971; Wickelgren, 1968).

Perhaps the most direct motivation for the present experiment comes from a recent study reported by Vandierendonck (1988), who also proposed the rule-instantiation model and tested its predictions against those of the exemplar model. Vandierendonck, who used a category structure similar to the present one, found that subjects' postacquisition typicality judgments were highly correlated with the distance of stimuli from the category boundary. Indeed, Vandierendonck's study *avored* the predictions of the rule model over those of the exemplar model, because slightly higher correlations were achieved with the former model as opposed to the latter. However, as Vandierendonck acknowledged (p. 73), in his experimental design the distance-from-boundary measure was itself highly correlated with various measures stemming from the exemplar-similarity hypothesis. Thus, his study did not provide for strong qualitative contrasts between the predictions of the competing models.

There are two main differences between the present design and the one used by Vandierendonck (1988). The most important difference is that in Vandierendonck's study, the stimuli that were most distant from the category boundary were always presented during training. By contrast, the present experiment included conditions in which the extreme stimuli were never presented during training. A second difference is that in the present experiment, the presentation frequencies of values on the irrelevant dimension were also manipulated across conditions. As will be seen, these manipulations yielded strong qualitative contrasts between the predictions stemming from the rule-instantiation and exemplar model hypotheses.

Theoretical Considerations Involving the Exemplar Model

Before we proceed to the report of the experiment, some additional theoretical considerations need to be discussed. The nature of the predictions for the exemplar-similarity model depend critically on two main factors: (1) the method by which similarities among exemplars are computed, and (2) the decision rule that translates exemplar similarities into typicality judgments.

On the similarity gradient. According to the exemplar model, the probability that stimulus i is classified into Category B is found by summing the similarity of stimulus i to all exemplars of Category B (weighted by their "strength" in memory), and then dividing by the summed

similarity of stimulus i to all exemplars of both categories:

$$P(R_B | S_i) = \frac{\sum_{b \in B} N_b s_{ib}}{\sum_{b \in B} N_b s_{ib} + \sum_{a \in A} N_a s_{ia}}, \quad (1)$$

where s_{ib} denotes the similarity between items i and b , and N_b denotes the strength with which exemplar b is stored in memory. In previous research (Estes, 1986a, 1986b; Nosofsky, 1988a, 1988b), the exemplar-strength terms were assumed to be directly proportional to the frequency with which the exemplars were presented during training. However, more recent research suggests that exemplar strength may be a negatively accelerated, increasing function of presentation frequency (Nosofsky, in press). Although not the focus of this article, other variables, such as recency of presentation, presumably also influence exemplar strength.

The similarity between exemplars i and b (s_{ib}) is assumed to be a decreasing function of their psychological distance (d_{ib}). Extensive research points to a similarity function of the form

$$s_{ib} = \exp(-d_{ib}^p). \quad (2)$$

The value of p in Equation 2 that provides the best description of the gradient relating similarity to psychological distance appears to depend on experimental conditions. An exponential decay function ($p = 1$) operates in "generalization" experiments involving highly discriminable stimuli and minimal training (Shepard, 1958a, 1987), whereas a Gaussian gradient ($p = 2$) operates in "perceptual discrimination" experiments involving confusable stimuli and extensive training (Nosofsky, 1985a, 1985b, 1986, 1989; Shepard, 1986). (For theoretical explanations of the changing similarity gradient, see Ennis, 1988; Ennis, Palen, & Mullen, 1988; and Shepard, 1958a, 1986, 1987.)

The form of the similarity gradient influences the exemplar model's predictions of classification performance for "rule-boundary" problems such as those illustrated in Figures 1 and 2. An illustration for a canonical case is given in Figure 3. The experiment requires the subject to discriminate between two stimuli lying on a unidimensional continuum, say size. Stimulus a is to be classified into Category A, and stimulus b into Category B. Stimuli a and b are separated by psychological distance d . Imag-

Figure 3. Schematic unidimensional category structure involving two training exemplars (a and b) separated by distance d , and one transfer stimulus (T) separated from instance b by distance x .

ine that after initial discrimination training on a and b transfer stimulus T is presented. As is illustrated in Figure 3, T is separated from the larger training stimulus by psychological distance x . It seems intuitively clear that T would be classified into Category B, because the "rule" that the subject has learned is that the small training stimulus is an A and the large one a B. Suppose, however, that classification decisions are based on exemplar-similarity comparisons. One might intuit that the exemplar model would predict roughly 50% classification responding, since T is highly dissimilar to both training exemplars. The actual predictions are otherwise.

First, consider an exponential similarity gradient. According to the exemplar model, the probability that T is classified into Category B is given by

$$\begin{aligned} P(R_B | T) &= \frac{s_{Tb}}{s_{Tb} + s_{Ta}} \\ &= \frac{\exp(-x)}{\exp(-x) + \exp[-(d+x)]} \\ &= \frac{1}{1 + \exp(-d)} \end{aligned} \quad (3)$$

Thus, the exponential-similarity exemplar model predicts that the probability with which T is classified into Category B is independent of its distance from the Category B training exemplar (x). In fact, the probability that transfer stimulus T is classified into Category B is equal to the probability with which training exemplar b is itself classified into Category B. Unfortunately, this counterintuitive prediction involving classification data is extremely difficult to test. The exponential gradient operates when the stimuli are readily discriminable. If a and b are highly discriminable, the probability with which both b and T are classified into Category B would be near unity. (It is important to note that if T were actually presented during classification training, the exponential-similarity exemplar model would predict that the most extreme classification probabilities would indeed be associated with T .)

Consider instead a Gaussian similarity gradient. According to the exemplar model, the probability that T is classified into Category B is given by

$$\begin{aligned} P(R_B | T) &= \frac{\exp(-x^2)}{\exp(-x^2) + \exp[-(d+x)^2]} \\ &= \frac{1}{1 + \exp[-(d^2 + 2dx)]} \end{aligned} \quad (4)$$

Now, as distance (x) between the transfer stimulus and the training exemplar increases, the predicted classification probability for the transfer stimulus approaches unity. That is, the larger the transfer stimulus, the higher the probability with which it is predicted to be classified in the "large" category, just as would be predicted by the distance-from-boundary model.

In summary, although transfer stimulus T is highly dissimilar to both training exemplars, the exemplar model

predicts that it will be classified into the "large" category with high probability. In this regard, the predictions of the exemplar and rule-based models are quite similar.

On the decision rule. Although the exemplar model and the rule model make similar predictions regarding classification probabilities, the models diverge in their predictions of typicality or goodness-of-example judgments. The reason is that, at least under certain conditions, different decision rules have been found to mediate classification and typicality judgments. According to the exemplar model, the probability with which a probe is classified into a target category is determined by the probe's similarity to the exemplars of the target category *relative* to the probe's similarity to the exemplars of contrast categories—see Equation 1. Throughout this article, I will refer to the decision rule embodied in Equation 1 as the *relative-similarity* measure. By contrast, an important determinant of typicality judgments often involves the *absolute* summed similarity of a probe to the exemplars of the target category (Nosofsky, 1988a). This *absolute-similarity* measure is given by the numerator in Equation 1. To the extent that absolute summed similarity to stored exemplars is involved in typicality judgments, the exemplar model will yield sharply contrasting predictions from the rule model. For example, consider Transfer Stimuli 15 and 16 in Figure 2. Although the exemplar model predicts that these large transfer stimuli will be classified into Category B with high probability, these transfer stimuli may not be rated as good members of Category B. The reason is that the absolute similarity of Transfer Stimuli 15 and 16 to the Category B exemplars is low.

To summarize, if people's category representations consist solely of the simple logical rule, with goodness judgments based on distance from the rule boundary (see, e.g., Vandierendonck, 1988), then the large Transfer Stimuli 15 and 16 should be judged as the best examples of Category B. By contrast, to the extent that people's category representations involve storage of individual exemplars and that goodness judgments are influenced by absolute summed similarity, the large transfer stimuli should have diminished goodness ratings. Beyond testing this qualitative contrast between the predictions of the rule and exemplar models, a goal in this research was to test the alternative models on quantitative grounds.

EXPERIMENT 1A

The experiment was organized into six conditions involving the category structure illustrated in Figure 2. All conditions consisted of a classification learning phase followed by a typicality-ratings transfer phase. The six conditions differed only in the relative frequency with which individual exemplars were presented during classification training. In Conditions HF10/ $\bar{L}$ and HF12/ $\bar{L}$, the large Transfer Stimuli 15 and 16 were never presented during training: Training Exemplars 10 and 12 were presented with high frequency in Conditions HF10 and HF12, re-

spectively. In Conditions HF10/L and HF12/L, the large Stimuli 15 and 16 were presented with the same frequency as were the other exemplars. Again, Exemplars 10 and 12 were presented with high frequency in Conditions HF10 and HF12, respectively. And in Conditions HF15 and HF16, Stimuli 15 and 16 were presented with high frequency, respectively. Thus, across conditions, there were manipulations of the frequency with which values on both the relevant and irrelevant dimensions were presented during training.

Method

Subjects

The subjects were 363 undergraduates from Indiana University who participated as part of an introductory psychology course requirement. There were 336 subjects in the classification learning experiment and 27 subjects in a preliminary similarity-scaling study. Of the 336 subjects in the classification-learning experiment, 36 were replacements for subjects who did not meet preestablished learning criteria. After replacement, there were 50 subjects in each of the six conditions of classification learning.

Stimuli and Apparatus

The stimuli were circles varying in size and angle of orientation of a radial line. Five levels of size and five levels of angle were combined orthogonally to yield a total of 25 stimuli. All 25 stimuli were used in the similarity-scaling study, and the subset of 16 stimuli in Figure 2 was used in the classification-learning experiment. The five levels of size (measured in terms of length of radial line) were 5.50, 6.25, 7.25, 8.00, and 19.00 mm; the five levels of angle of radial line (measured counterclockwise from horizontal) were 5°, 30°, 45°, 60°, and 85°. The stimuli were presented on VR14 scopes; they appeared as continuous green images on a black background. The experiment was controlled by a PDP-11/34 computer. The subjects entered responses by using response boxes that were interfaced with the computer.

Procedure

Classification learning. To facilitate the description of the procedure and the reporting of the results, the stimuli will often be referenced with their level of size. For example, Stimuli 1-3 will be referred to as the Size 1 stimuli, and Stimuli 15 and 16 as the Size 5 stimuli. (The term *Size 4 stimuli* will be reserved for Stimuli 10-12, whereas Stimuli 13 and 14, which were never presented during classification training in any conditions, will be referred to as the *Size 4* stimuli*—see Figure 2.)

On any given trial, the stimulus was presented on the screen, and the subject judged whether it belonged to Category A or to Category B. Feedback was provided on each trial, following the response.

There were six conditions, with each condition differing only in the frequency with which individual stimuli were presented during classification training. In each condition, a single stimulus was presented with high frequency. In Conditions HF10/ $\bar{L}$ and HF12/ $\bar{L}$ (referred to collectively as Condition $\bar{L}$), the Size 5 stimuli were never presented during training. In Condition HF10/ $\bar{L}$ (respectively, HF12/ $\bar{L}$), Stimulus 10 (respectively, 12) was presented with probability .45, and the remainder of Stimuli 1-12 were presented with probability .05 each. In Conditions HF10/L and HF12/L (collectively, Condition L), the Size 5 stimuli were presented with the same frequency as were the other stimuli, with the exception of the single high-frequency stimulus. In Condition HF10/L (respectively, HF12/L), Stimulus 10 (respectively, 12) was presented with probability .443 and the remainder of Stimuli 1-12, 15, and 16 with probability .043 each. Finally, in Conditions HF15 and HF16 (col-

lectively, Condition HFL). Stimuli 15 and 16, respectively, were presented with probability .443, and the remaining stimuli with probability .043 each. Stimulus presentations were random within the constraints stated above. There was a total of 100 trials of training in Conditions HF10/L and HF12/L, and 110 trials of training in the remaining four conditions. There were slightly more trials in the latter conditions in order to roughly equate the absolute frequency with which most individual exemplars were presented during training across conditions.

Transfer. Following classification training, the subjects engaged in a typicality-judgment transfer phase. Two tests were administered in this phase. First, the subjects engaged in typicality paired-comparison judgments. On each trial, a pair of stimuli was presented, and the subjects judged which stimulus was a better example of Category B. In all conditions, the first 15 trials were paired comparisons involving Stimuli 7, 9, 10, 12, 13, and 14 (see Figure 2). All possible pairs of these six stimuli were presented, with order randomized for each subject. In Conditions HF10/L, HF10/L, and HF15, the 16th trial was a paired comparison involving Stimuli 10 and 15, whereas in Conditions HF12/L, HF12/L, and HF16, the 16th paired comparison consisted of Stimuli 12 and 16. Note that the latter paired comparisons pitted a high-frequency exemplar against an exemplar that maximally instantiated the small-large rule. Finally, in Conditions HF15 and HF16, there was a 17th paired comparison involving Stimuli 15 and 16. Left-right placement of the stimuli within each pair was randomized on each trial.

The second transfer test consisted of direct typicality ratings. On each trial, one of the 16 stimuli was presented, and subjects rated how good an example it was of Category B. The ratings ranged from 1 (very poor example) to 10 (excellent example), and the subjects were urged to use the full range of ratings. Order of presentation of the stimuli was randomized for each subject.

Similarity-scaling study. In addition to the classification learning experiment, a similarity-scaling condition was tested. Similarities between exemplars were measured in terms of confusions in an identification experiment. On each trial, one of the 25 stimuli was randomly selected and presented on the VR14 scope, and the subject identified its size and angle. Values on the size and angle dimensions were each labeled with the numbers 1 (smallest size, lowest angle) through 5 (largest size, highest angle). To enter a response of size *i*-angle *j*, the subject first pressed button *i* and then pressed button *j*. Following the responses, corrective feedback was provided. There were 150 trials in the identification condition.

Results and Discussion

Analysis of the postacquisition typicality ratings and paired-comparisons data is restricted to the subjects who met a preestablished learning criterion during the training phase. The criterion was that subjects respond with greater than 50% accuracy on each of the individual Size 1 exemplars (1, 2, and 3) and Size 4 exemplars (10, 11, and 12) during the final 60 trials of classification training. The reason for establishing this criterion was the assumption that any subject who failed to respond accurately to these stimuli had not "learned the rule" for the category (i.e., the small-large distinction). It was thought that including "non-rule-learners" in the analyses would be unfair to testing the predictions of the rule-based model. Use of this learning criterion led to replacing an average of 6 subjects in each condition, ranging from 1 subject in Condition HF10/L to 9 subjects in Condition HF16.

Table 1
Mean Category B Typicality Ratings for Each Stimulus as a Function of Conditions in Experiment 1A

Size Level	Stimulus	Condition					
		HF10/L	HF12/L	HF10/L	HF12/L	HF15	HF16
1	1	2.76	2.12	1.62	1.66	1.74	1.40
	2	2.86	2.22	1.66	1.62	2.06	1.74
	3	2.58	2.12	1.64	1.60	1.60	2.08
2	4	3.96	3.42	2.62	2.98	2.58	2.96
	5	4.30	4.02	2.72	3.28	3.24	2.48
	6	3.94	3.42	2.66	3.06	2.50	2.88
3	7	7.54	7.16	6.78	7.16	5.94	5.76
	8	7.30	7.12	7.02	7.24	5.76	5.90
	9	7.60	7.36	6.80	6.98	6.16	5.98
4	10	8.34	8.02	7.74	8.10	6.68	6.30
	11	8.72	8.80	8.02	8.20	7.00	6.82
	12	8.42	8.82	7.84	8.20	6.44	6.26
4*	13	6.56	6.18	5.68	5.30	4.70	4.22
	14	6.62	6.40	5.82	5.26	4.12	4.22
5	15	4.58	4.06	9.08	8.26	9.46	9.06
	16	4.36	4.14	8.98	8.28	9.42	9.46

Relevant Dimension (Size)

The mean Category B typicality ratings for each stimulus in each condition are shown in Table 1. If we restrict our attention first to Stimuli 1-12, which were all presented during classification training in all conditions, we can see that typicality ratings increased as the size of the stimuli increased. This result is as expected and is predicted by both the rule and the exemplar models (see the Theoretical Analyses section).

More interesting is that the frequency manipulations of values on the relevant dimension had a major impact on the typicality ratings. The most important effect is illustrated in Figure 4, in which the mean typicality ratings for the Size 4 stimuli (10, 11, and 12) and for the Size 5 stimuli (15 and 16) are plotted for each condition. In the figure, the data are averaged over conditions in which the frequency of values on the irrelevant dimension was manipulated; thus, for example, for Condition L, the data are averaged over Conditions HF10/L and HF12/L. (Such averaging is justified by the observation that the frequency manipulations on the irrelevant angle dimension had little effect on the typicality ratings.) In Condition L, in which the Size 5 stimuli were never presented during training, the Size 4 stimuli were rated as the best examples of Category B, whereas the Size 5 stimuli were rated as relatively poor examples. However, as the Size 5 stimuli were presented with higher frequency during training, their typicality ratings increased dramatically and exceeded the typicality ratings for the Size 4 stimuli in Conditions L and HFL. A 3×2 analysis of variance with conditions (L, L, and HFL) and stimuli (Size 4 and Size 5) as factors revealed a highly significant interaction [$F(2,297) = 140.3$, $MS_e = 4.59$, $p < .001$]. In *t* tests, it was confirmed that the mean typicality ratings

Figure 4. Mean Category B typicality ratings for the Size 4 stimuli (solid circles) and Size 5 stimuli (open circles) as a function of conditions in Experiment 1A. In Condition $\bar{L}$, the large Size 5 stimuli were never presented during training, whereas in Conditions L and HFL, the Size 5 stimuli were presented with normal and high frequency, respectively.

for the Size 4 stimuli exceeded those of the Size 5 stimuli in Condition $\bar{L}$ [$t(99) = 10.41, p < .001$], but the reverse was observed in Conditions L [$t(99) = -2.57, p < .01$] and HFL [$t(99) = -12.49, p < .001$].

The typicality paired-comparison data are presented in Appendix A. The paired-comparison results mirrored the patterns observed for the typicality ratings. Across all conditions, the Size 4 stimuli (10 and 12) were judged as better examples of Category B than were the Size 3 stimuli (7 and 9). The critical finding concerns the comparisons for the Size 4 and Size 5 stimuli, which are summarized in Figure 5. The figure shows, for each condition, the average probability with which either the Size 4 or the Size 5 stimuli were judged the better examples of Category B. (The data are collapsed across conditions in which values on the irrelevant angle dimension were manipulated. Note that *within* each condition, the Size 4 and Size 5 stimuli that were compared had identical values of angle.) In Condition $\bar{L}$, in which the Size 5 stimuli were never presented during training, the Size 4 stimuli were judged as better examples of Category B than were the Size 5 stimuli, but the reverse occurred in Conditions L and HFL. Preference for the Size 5 stimuli increased dramatically as their frequency of presentation was increased. A chi-square test of independence revealed that this interaction between size preference and learning conditions was highly significant [$\chi^2(2) = 90.8, p < .0001$].

The results for both the typicality ratings and the typicality paired comparisons are inconsistent with the idea that people's mental representation of the category consisted solely of a boundary partitioning small and large stimuli. The most natural prediction stemming from such a model is that the larger the stimulus, the better an example it should be of Category B. However, in Condition $\bar{L}$, in which the Size 5 stimuli were never presented during training, the Size 4 stimuli were rated as far bet-

ter examples of the category. The result is consistent with a model in which it is assumed that individual exemplar traces form an important part of the category representation.

In Condition L, the Size 5 stimuli were rated as better examples of Category B than were the Size 4 stimuli, despite the Size 4 stimuli's having been presented with high frequency during training. This result is inconsistent with the idea that people's typicality judgments were based solely on absolute summed similarity to stored exemplars of the target category. The extent to which the results can be explained in terms of a mixture of the absolute-similarity and relative-similarity measures is explored in the Theoretical Analyses section.

Irrelevant Dimension (Angle)

With the possible exception of Stimuli 13 and 14, the frequency manipulation of values on the irrelevant dimension had only small and insignificant effects on people's typicality ratings. However, clear effects emerged in the typicality paired-comparison data. Figure 6 shows the probability with which individual stimuli in critical test pairs were judged as better examples of Category B as a function of whether Angle 2 or Angle 4 was presented with high frequency during training. In each panel of Figure 6, the frequency of values on the relevant size dimension is held constant. The solid line in each panel corresponds to stimuli that received direct frequency manipulations (i.e., 10 and 12 in Conditions $\bar{L}$ and L, and 15 and 16 in Condition HFL), and the dashed line corresponds to generalization items (i.e., 7 and 9 in Conditions $\bar{L}$ and L, 10 and 12 in Condition HFL).

The frequency manipulation on the irrelevant angle dimension exerted a powerful influence on people's paired-comparison goodness-of-example judgments in Conditions

Figure 5. Mean probability with which either the Size 4 stimuli (solid circles) or the Size 5 stimuli (open circles) were judged the better examples of Category B, as a function of conditions in Experiment 1A. In Condition $\bar{L}$, the large Size 5 stimuli were never presented during training, whereas in Conditions L and HFL, the Size 5 stimuli were presented with normal and high frequency, respectively.

Figure 6. Probability with which individual stimuli in critical test pairs were judged the better example of Category B as a function of whether Angle 2 or Angle 4 was presented with high frequency during training (Experiment 1A).

L and HFL.¹ For example, in Condition HF12/L, Stimulus 12 was judged a better example of Category B than was Stimulus 10 (with probability .72), but in Condition HF10/L, Stimulus 10 was preferred to Stimulus 12 (with probability .60). These crossovers in paired-comparison judgments also existed for generalization items. For example, in Condition HF12/L, Stimulus 9, which was highly similar to Stimulus 12 and had the same value of angle, was preferred to Stimulus 7 (with probability .62), but in Condition HF10/L, Stimulus 7 was preferred to Stimulus 9 (with probability .58). Chi-square tests of independence revealed that these crossover effects were significant for Pairs 10-12 and 7-9 in Condition L, and for Pairs 15-16 and 10-12 in Condition HFL [average $\chi^2(1) = 8.55$, $p < .05$].

Stimuli 13 and 14, which had highly distinctive angle values that were never presented during training, received low typicality ratings and paired-comparison preference probabilities in all conditions. Strong conclusions cannot be drawn regarding the typicality results for Stimuli 13 and 14, however, because the frequency with which Angle Values 1 and 5 were presented was not manipulated in the experiment. However, it is plausible that the low typicality ratings associated with these stimuli resulted, at least in part, because they were not previously experienced and were highly distinctive from the original training exemplars.

The finding that frequency of values on the irrelevant angle dimension influenced people's paired-comparison typicality judgments is inconsistent with the rule-based model. Because stimuli of the same size but with different angles are the same distance from the category boundary, they should show equal category goodness. The results cannot be explained by positing that the different angle values influenced the size judgments (e.g., that

stimuli with Angle 2 appeared larger than stimuli with Angle 4), because the typicality judgments changed systematically depending on which angle value was presented with high frequency. The typicality data are consistent with the idea that individual exemplar traces form part of the category representation and that at least some storage of values on the irrelevant angle dimension took place.

Theoretical Analyses

This section summarizes theoretical analyses in which the exemplar model was tested on its ability to predict quantitatively the typicality data. Details of the theoretical analyses are presented in Appendix B. The first step was to analyze the data obtained in the preliminary identification condition (see the Method section) to derive a multidimensional scaling (MDS) solution for the exemplars and to determine the nature of the similarity gradient relating similarity to distance in the psychological space (e.g., Nosofsky, 1985b; Shepard, 1958b, 1987). The derived MDS solution is shown in Figure 2, with the analyses having supported an exponential decay similarity gradient (Shepard, 1987). The MDS solution and exponential similarity gradient were then used in conjunction with the exemplar model to predict the typicality data (cf. Nosofsky, 1986, 1987, 1989).

The upshot of the theoretical analyses was that a mixture of the relative-similarity and absolute-similarity measures discussed previously in this article was necessary to predict the Category B typicality ratings and paired comparisons. The relative-similarity measure for an item was based on the context model's prediction of the Category B response probability for that item (Equation 1). The absolute-similarity measure was obtained by summing the

similarity of an item to all exemplars of Category B. An item's "typicality strength" was then assumed to be based on a mixture of these measures. The subjects' typicality ratings were assumed to be ordinally related to the derived typicality strengths. The predicted probability that one stimulus was selected as more typical than another in the paired-comparison judgments was given by a logistic transformation of the differences in typicality strengths of the items. The model required estimation of four parameters to predict 96 typicality ratings, and five parameters to predict 98 paired comparisons. For the 96 typicality ratings, the mixed model achieved a Spearman rank-order correlation of .974. Also, the mixed model accounted for 90.3% of the variance in the paired-comparisons data. Quantitative versions of the rule-instantiation model as well as pure relative-similarity and absolute-similarity versions of the exemplar model failed dramatically to predict the data. The rule-instantiation model and the pure relative-similarity exemplar model incorrectly predicted high typicality for the Size 5 stimuli in Conditions HF10/ $\bar{L}$ and HF12/ $\bar{L}$ and for Stimuli 13 and 14 in all conditions. The pure absolute-similarity exemplar model incorrectly predicted low typicality ratings for the Size 5 stimuli in Conditions HF10/L and HF12/L and also incorrectly predicted that in Conditions HF15 and HF16, typicality ratings for the Size 3 and Size 4 stimuli would be virtually identical.

Although the mixed model provided a good first approximation to the typicality data, it could be easily rejected on quantitative grounds (see Appendix B). Because typicality judgments are highly open-ended measures and undoubtedly involve complex cognitive processes, the quantitative failing of the model is not surprising. Perhaps of more concern is the interpretation of the mixed model: Does it imply a mixing of relative similarity and absolute similarity at the individual subject level, or is it primarily reflecting a between-subject phenomenon? A partial answer to this difficult question is provided by consideration of the variability estimates of typicality ratings for the Size 5 stimuli in Condition $\bar{L}$. The between-subjects variance for the Size 5 stimuli was more than twice as large as for any of the remaining stimuli (some of which had roughly the same mean typicality ratings). In other words, many subjects gave fairly high typicality ratings to the Size 5 stimuli in Condition $\bar{L}$, apparently emphasizing the relative-similarity component, whereas many other subjects gave fairly low ratings to these stimuli, apparently emphasizing the absolute-similarity component. (The relative weighting of the two similarity components is not predicted by the exemplar model to exert as much impact on the ratings for the remaining stimuli as for the Size 5 stimuli.) This diagnostic suggests a major contribution of mixing at the between-subjects level, although mixing at the within-subjects level may also, of course, have occurred.

Admittedly, there is an inelegance associated with the need to use the mixed model to predict the typicality data. Nevertheless, the result is consistent with much recent work involving typicality judgments. For example, in a

study of the determinants of graded structure in various types of natural categories, Barsalou (1985, p. 647) concluded: "In general, it appears that there are many different kinds of information available for people to use when judging typicality ... [and] it appears that people may often simultaneously incorporate several kinds of information into their judgments of typicality." Likewise, in studies of artificial category learning, Nosofsky (1988b) observed a close correspondence between typicality judgments and classification probabilities, but also noted the clear divergence in these measures that was reported by Bourne (1982). Nosofsky (1988b) suggested: "'Typicality' is an open-ended construct, and the interpretation given to it by people apparently may vary depending on instructions and experimental conditions" (p. 63). Apparently, the present experimental conditions led subjects to use at least two kinds of information when they made their typicality judgments—what are interpreted here as relative and as absolute summed similarity.

EXPERIMENT 1B

A theorist wishing to maintain the primacy of the logical-rule component in subjects' category representations could argue that the rule boundary is formed and used for making classification decisions, but that exemplars are also stored as part of the learning process. Furthermore, in making goodness judgments, the subject emphasizes the exemplar memories, despite the fact that the rule is used for classification. Although it is unclear why distance from the rule boundary would not govern goodness-of-example judgments if the rule component were indeed primary, the hypothesis cannot be dismissed. The purpose in Experiment 1B was to demonstrate a limitation of the distance-from-boundary model as the sole basis for *actual classification decisions*, and to suggest a role of stored exemplars even here. As noted in the introduction, for the case of the present logically defined category structure, the predictions of the rule model and the exemplar model are extremely difficult to disentangle on the basis of classification probability data. In Experiment 1B, the attempt was made to contrast the models on the basis of reaction times collected in a speeded classification task.

The learning phases of Conditions HF12/ $\bar{L}$, HF12/L, and HF16 were replicated and followed by speeded classifications. The 16 transfer stimuli were presented in a random order, and subjects were asked to classify them into Category A or B as quickly as possible without making errors. Based on the proposals of numerous researchers, the predictions of the distance-from-boundary model are clearcut: The larger the transfer stimulus, the more quickly it should be classified in Category B (e.g., Ashby & Maddox, in press; Cartwright, 1941; Emmerich et al., 1972; Hyman & Frost, 1975; Pike, 1973). We are on shakier ground with regard to the exemplar model, because an explicit process model of classification reaction time has not been proposed within this framework. To the extent that absolute summed similarity to stored ex-

emplars is involved in speeded classification, however, we expect increased reaction times for the very large transfer stimuli that are not yet stored as part of subjects' category representations.

The present experiment is similar to a previous one reported by Hyman and Frost (1975), who contrasted a rule-instantiation model and an exemplar model on their predictions of speeded classifications. One difference between the studies is that Hyman and Frost used random "dot pattern" stimuli that were systematically varied in their height and width. Although a two-dimensional scaling solution for the patterns reflected these variations in height and width, Hyman and Frost reported that various configural properties of the dot patterns apparently also influenced the speeded classifications. The highly simplified two-dimensional stimuli used in the present study may allow for a purer test between the rule and exemplar models. Another difference between the studies is that Hyman and Frost did not manipulate the frequency of the patterns that were most distant from the category boundary, as was done here. These frequency manipulations allowed for strong qualitative contrasts between the predictions of the competing models, which were not available in the Hyman and Frost (1975) study.

Method

Subjects. The subjects were 159 undergraduates from Indiana University who participated as part of a course requirement. After replacing 24 subjects who failed to meet the learning criterion, there were totals of 53, 40, and 42 subjects in Conditions HF12/L, HF12/L, and HF16, respectively.

Stimuli and Apparatus. The stimuli and apparatus were the same as in Experiment 1A.

Procedure. The procedure was the same as in Conditions HF12/L, HF12/L, and HF16 of Experiment 1A, except that the subjects engaged in a speeded classification task during the transfer phase. There were two blocks of speeded classifications. In each block, the 16 transfer stimuli were presented in a random order. At the start of each trial, a fixation point appeared on the screen for 1 sec. The transfer stimulus was then presented immediately, and it remained on the screen until the subject entered a response. The subjects rested the left index finger on the Category A button and the right index finger on the Category B button throughout the testing session. They were instructed to classify the stimuli into either Category A or Category B as quickly as possible without making errors.

Results and Discussion

In the summarizing of the reaction time (RT) data and in the analyses, all individual RTs of greater than 1,500 msec were deleted. Such RTs were rare (approximately 2% of the data), they were greater than two standard deviations from the mean for all subjects.

The mean correct RTs for each stimulus are presented separately for Blocks 1 and 2 in Tables 2 and 3. Also shown are the error frequencies for each stimulus. Note that errors were rare for the Size 4 and Size 5 stimuli, which are the main items of interest.

The critical results are shown in Figure 7, in which mean RT for the Size 4 stimuli and the Size 5 stimuli in Block 1 is plotted as a function of conditions. The frequency manipulation exerted a dramatic impact on the pat-

Table 2
Mean Correct Reaction Times (RT) and Error Frequencies (EF) for Each Stimulus as a Function of Conditions in Block 1 of Experiment 1B

Size Level	Stimulus	Condition					
		HF12/L		HF12/L		HF16	
		RT	EF	RT	EF	RT	EF
1	1	501	2	534	1	484	3
	2	514	1	497	5	482	2
	3	465	2	502	2	510	1
2	4	510	12	633	11	642	15
	5	525	12	640	10	572	12
	6	613	12	569	6	601	11
3	7	508	9	513	1	510	2
	8	495	8	460	1	583	2
	9	570	5	497	3	538	2
4	10	465	3	428	2	472	1
	11	480	1	448	0	511	0
	12	469	3	464	0	473	0
4*	13	506	4	511	1	499	0
	14	583	3	475	1	539	1
5	15	520	1	471	0	430	0
	16	518	1	474	0	434	0

Note—Maximum error frequencies for each stimulus were 53, 40, and 42 in Conditions L, L, and HFL, respectively.

Table 3
Mean Correct Reaction Times (RT) and Error Frequencies (EF) for Each Stimulus as a Function of Conditions in Block 2 of Experiment 1B

Size Level	Stimulus	Condition					
		HF12/L		HF12/L		HF16	
		RT	EF	RT	EF	RT	EF
1	1	484	2	500	1	510	2
	2	477	0	452	0	487	1
	3	466	1	515	4	481	0
2	4	610	6	597	8	584	15
	5	493	5	600	4	626	11
	6	547	7	625	6	565	11
3	7	512	8	514	3	524	1
	8	516	10	526	1	524	3
	9	500	4	513	1	495	4
4	10	504	0	447	2	481	1
	11	480	3	471	0	453	0
	12	468	1	496	2	476	1
4*	13	483	1	482	1	498	1
	14	527	3	517	1	480	1
5	15	465	1	448	0	427	0
	16	457	1	435	0	398	0

Note—Maximum error frequencies for each stimulus were 53, 40, and 42 in Conditions L, L, and HFL, respectively.

tern of RTs. In Condition HF12/L, in which the Size 5 stimuli were never presented during training, the Size 4 stimuli were classified significantly faster than the Size 5 stimuli [$t(53) = -2.32, p < .05$]. Impressively (and unexpectedly), the same pattern of results was observed in Condition HF12/L, in which the Size 5 stimuli were presented with low frequency during training. The difference in RTs between the Size 4 and Size 5 stimuli was only

Figure 7. Mean correct reaction times for the Size 4 stimuli (solid circles) and the Size 5 stimuli (open circles) as a function of conditions in Block 1 of Experiment 1B. In Condition $\bar{L}$, the large Size 5 stimuli were never presented during training, whereas in Conditions L and HFL, the Size 5 stimuli were presented with normal and high frequency, respectively.

marginally significant in this condition [$t(40) = -1.83$, $p < .10$]. A t test done with means of log-transformed correct RTs, however, did yield statistical significance [$t(40) = -2.13$, $p < .05$]. Finally, in Condition HF16, the ordering of RTs was reversed from the ordering in the other conditions, with the Size 5 stimuli being classified significantly faster than the Size 4 stimuli [$t(42) = 2.60$, $p < .05$]. A 3×2 analysis of variance with conditions (HF12/ $\bar{L}$, HF12/L, HF16) and stimuli (Size 4, Size 5) as factors yielded a significant interaction [$F(2,132) = 6.79$, $MS_e = 10,711.4$, $p < .01$]. A 2×2 analysis of variance done with only the data from conditions in which the Size 5 stimuli were presented during training (HF12/L, HF16) also yielded a significant interaction [$F(1,80) = 9.73$, $MS_e = 8,283.6$, $p < .01$]. These interactions reflect the reversed orderings of RTs for the Size 4 and Size 5 stimuli across the first two conditions and Condition HF16.

The pattern of results in Block 1 argues against the idea that RT is simply a decreasing function of distance of an object from the category rule boundary. One might try to salvage the distance-from-boundary hypothesis by arguing that in Condition HF12/ $\bar{L}$, the presentations of the novel Size 5 stimuli in the transfer phase were "surprising." But in a sense, this argument is part and parcel of the exemplar model hypothesis. The Size 5 stimuli are "surprising" because they are not similar to any exemplars that are part of subjects' category representations. Furthermore, even in Condition HF12/L, in which the Size 5 stimuli had been previously experienced and would not have a surprise effect, there was an advantage for the high-frequency Size 4 stimuli. The most important result is simply that the pattern of RTs changed systematically as a function of the frequency manipulation. These results support the hypothesis that stored exemplars play a fundamental role in speeded classification for logically defined concepts.²

The RT advantage for the Size 4 stimuli over the Size 5 stimuli in Block 1 of Conditions HF12/ $\bar{L}$ and HF12/L disappeared by Block 2 (see Tables 2 and 3). This result was not unanticipated. Once the Size 5 stimuli have been viewed and classified, they may effectively become a strong component of subjects' representations of Category B (e.g., see Posner & Keele, 1968, and Nosofsky, 1986, for discussions of how category representations may evolve during transfer). Even in Block 2 of the transfer phase, however, there was a residual effect of the frequency manipulation. The mean RT differences between the Size 4 stimuli and the Size 5 stimuli during Block 2 were 23.0, 29.8, and 57.5 msec in Conditions HF12/ $\bar{L}$, HF12/L, and HF16, respectively. Thus, as relative presentation frequency for the Size 5 stimuli increased during training, they obtained a larger RT advantage during Block 2 of transfer.

In the rule model that was contrasted with the exemplar model in Experiments 1A and 1B, it was assumed that a single category boundary was established to partition the members of Categories A and B, as in traditional signal detection models of unidimensional classification. To explain the low goodness ratings and high RTs associated with the Size 5 stimuli in Condition $\bar{L}$, a rule-boundary theorist might posit that multiple boundaries were formed. In addition to the primary boundary partitioning the Category A and B exemplars, a second boundary was established between the Size 4 and Size 5 items. Unfortunately, this post hoc proposal fails to directly account for the finding that subjects in Experiment 1B essentially always classified the Size 5 items into Category B. Ashby and Gott (1988) and Ashby and Maddox (1990), for example, who have proposed and tested multidimensional signal detection models, have assumed that any percept falling in a region not demarcated by decision boundaries would result in a random response. Alternatively, one could argue that such ambiguous percepts are classified into the closest demarcated region, but this similarity-based proposal seems to add little explanatory power over the simple exemplar model. The multiple-boundary model also fails to account for the systematic effects of exemplar frequency on the typicality judgments and classification RTs across Conditions L and HFL, in which the Size 5 stimuli *had* been presented during training.

EXPERIMENT 2

The purpose in Experiment 2 was to use a new logically defined category structure to contrast the predictions of the rule-instantiation and exemplar-similarity models. In this experiment, members of a positive concept were defined by a disjunctive rule. The stimuli were geometric forms varying along four binary-valued dimensions: shape (squares or ellipses), size (large or small), position on screen (left or right), and number of contours (one or two). Two dimensions, randomly selected for each subject, were relevant to defining the rule. Let x and y denote the positive values on the first and second relevant dimen-

sions, respectively. Likewise, let $\bar{x}$ and $\bar{y}$ denote the negative values on these dimensions. A stimulus was defined to be a member of the positive concept if it had value x or y or both—membership was defined by an *inclusive-or* rule. The experimental manipulation was that across conditions, the presentation frequency of the stimulus with both positive values (xy) was varied. In the main condition of interest, xy was never presented during training. Now, to the extent that subjects learned and represented the positive concept solely in terms of the simple disjunctive rule, stimulus xy should be judged as the best example of the positive concept. For even though stimuli $x\bar{y}$ and $\bar{x}y$ also satisfy the disjunctive rule, the degree to which the rule is instantiated is greatest for stimulus xy . This rule-instantiation hypothesis is also illustrated in Figure 8, which shows that stimulus xy is the greatest distance from two plausible category boundaries that partition the members of the positive and negative concepts. In contrast to the rule-instantiation hypothesis, the exemplar model predicts that to the extent that absolute summed similarity is involved in typicality judgments, stimulus xy will have diminished goodness ratings.

The present experiment was related to Bourne's (1982) well-known study, which was seminal in demonstrating the existence of typicality gradients in logically defined categories. As in the present experiment, Bourne (1982) used a disjunctive concept and manipulated the presentation frequency of the xy stimulus. The main difference between Bourne's study and the present one is that in Bourne's study the xy stimulus was assigned probabilistically to both the positive *and* the negative concepts across conditions. In particular, in the condition in which xy was

never presented as a member of the positive concept, it appeared frequently as a member of the negative concept. In other words, in this condition of Bourne's (1982) experiment, the positive concept was defined by an *exclusive-or* rule—namely, x or y but not both. Under such conditions, there is little reason to expect xy to be judged as a good example of the positive concept, either on the basis of a rule-instantiation hypothesis or on that of an exemplar-similarity hypothesis. By contrast, in the present experiment, subjects were free to adopt the simple inclusive-or rule in all conditions. Note that on grounds of simplicity and cognitive economy, the inclusive-or rule, $x \vee y$, should be favored over the exclusive-or rule, $(x \wedge \bar{y}) \vee (\bar{x} \wedge y)$. Thus, the present design yielded sharply contrasting predictions from the rule-instantiation and exemplar-similarity hypotheses.

Although not the focus of the present research, the design of Experiment 2 also effectively contrasted the predictions of the exemplar model with those of independent feature-frequency models (e.g., Estes, 1986a; Kellogg, 1981; Reed, 1972). In the present design (see the Method section), independent feature-frequency models were constrained to predict that xy would be judged as at least as good a member of the positive concept as were $x\bar{y}$ and $\bar{x}y$, even in the condition in which xy was never presented during training. The reason is that the individual features that compose xy occur with the same frequency in the positive concept as do the individual features that compose $x\bar{y}$ (or $\bar{x}y$). Furthermore, the individual features that compose $x\bar{y}$ (or $\bar{x}y$) occur with greater frequency in the negative concept (the contrast category) than do the features composing xy . Unlike independent feature-frequency models, in the exemplar model it is assumed that people are sensitive to joint probabilities of features, and the model is therefore able to predict that xy will have diminished goodness ratings relative to $x\bar{y}$ and $\bar{x}y$. It is important that the inclusive-or concept-learning problem is solvable through an independent feature-frequency strategy. Therefore, the present experiment taps directly into the issue of people's preferences and biases for using independent-cue versus relational-cue coding strategies in classification learning (cf. Hayes-Roth & Hayes-Roth, 1977; Medin & Schaffer, 1978; Neumann, 1974; Reitman & Bower, 1973).

Method

Subjects

The subjects were 91 undergraduates from Indiana University who participated as part of a course requirement. After 1 subject who failed to meet a learning criterion had been replaced, there was a total of 30 subjects in each of three conditions of concept learning.

Stimuli and Apparatus

The stimuli were geometric forms varying along four binary-valued dimensions: shape (squares or ellipses), size (large or small), position (left or right), and number of contours (one or two). The stimuli were generated on an IBM-XT computer, which was used for controlling the experiment. Assignment of dimensions to the logical category structure was randomized for each subject, as was

Figure 8. Disjunctive-rule category structure. The dashed line represents a linear boundary for partitioning members of the positive and negative concepts, whereas the dotted line represents an independent-decisions boundary (see Ashby & Gott, 1988).

the choice of which values on the relevant dimensions were defined as positive values.

Procedure

Learning. Only any given trial, a stimulus was presented on the screen, and the subject judged whether or not it belonged to the positive concept. Feedback was provided on each trial, following the response. Learning continued for a maximum of 300 trials, or until the subject met a criterion of 32 consecutive correct responses.

There were three conditions. In all conditions, positive and negative concept trials occurred with equal probability—namely, .50. The conditions differed only in the prior probability with which stimulus xy was presented on positive concept trials. Across the three conditions, the probabilities that stimulus xy would be presented, given a positive concept trial, were .00, .10, and .33, respectively. On positive concept trials in which stimulus xy was not presented, stimuli $x\bar{y}$ and $\bar{x}y$ were presented with equal probability. Thus, across the three conditions, the probabilities that each of stimuli $x\bar{y}$ and $\bar{x}y$ would be presented, given a positive concept trial, were .50, .45, and .33, respectively. Of course, on negative concept trials, stimulus $\bar{x}\bar{y}$ was always presented. For both the positive and the negative concepts, the values presented on the two irrelevant dimensions were randomly selected on each trial. Stimulus presentations were randomized within the constraints stated above, except that in the conditions in which stimulus xy was presented with nonzero probability, there was at least one occurrence of xy during Trials 20–32 of a subject's sequence of consecutive correct responses.

Transfer. Following acquisition, there was a transfer phase in which subjects made paired-comparison goodness-of-example judgments and then direct typicality ratings. In the paired-comparison task, a pair of stimuli was presented on the screen (one at the top and one at the bottom), and subjects judged which was a better example of the positive concept. The types of paired comparisons were $x\bar{y}$ - xy , $\bar{x}y$ - xy , $x\bar{y}$ - $\bar{x}y$, $\bar{x}y$ - $\bar{x}\bar{y}$, and xy - $\bar{x}\bar{y}$. Each type of paired comparison occurred three times for each subject, for a total of 15 paired comparisons per subject. Order of presentation of the paired comparisons was randomized, as were the values selected for the irrelevant dimensions.

In the typicality-ratings phase, the subjects were presented with each of the 16 possible stimuli generated from the four binary-valued dimensions, and they rated on a scale from 1 (very poor) to 10 (excellent) how good an example the stimulus was of the positive concept. Note that there were four examples of each of the four types of stimuli, xy , $x\bar{y}$, $\bar{x}y$, and $\bar{x}\bar{y}$. Order of presentation of the 16 stimuli was randomized for each subject.

Results and Discussion

Learning

In the report of the results that follows, the three conditions will be denoted by the probability with which xy was presented during training on positive concept trials, namely, P_{00} , P_{10} , and P_{33} . The average numbers of trials to completing the criterion run were 80.9, 87.8, and 106.7, in Conditions P_{00} , P_{10} , and P_{33} , respectively. The differences among the groups were not statistically significant [$F(2,87) = 2.11$, $MS_e = 2,532.9$, $p > .05$].

Paired-Comparison Judgments

The results for the paired-comparison judgments are shown in Table 4. The main result of interest concerns the $x\bar{y}$ - xy and $\bar{x}y$ - xy paired comparisons. Stimuli $x\bar{y}$ and $\bar{x}y$ were clearly preferred to stimulus xy in Condition P_{00} , but the reverse was observed in Conditions P_{10} and P_{33} . Thus, the frequency manipulation had a major impact on

Table 4
Paired-Comparison Preference Probabilities in Experiment 2

Pair Type	Condition					
	P_{00}		P_{10}		P_{33}	
	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.
$x\bar{y}$ vs. xy	.69	.64	.39	.50	.23	.22
$\bar{x}y$ vs. xy	.68	.64	.40	.50	.24	.22
$x\bar{y}$ vs. $\bar{x}y$	.94	.92	.96	.92	.91	.92
$\bar{x}y$ vs. $\bar{x}\bar{y}$	.91	.92	.96	.92	.92	.92
xy vs. $\bar{x}\bar{y}$	.78	.86	1.00	.92	.96	.98

Note—Obs. = observed preference probability. Pred. = predicted preference probability. Table gives the probability with which the first member in each type of pair was judged more typical than the second member in each of the three conditions. Each probability is based on 90 observations. Also given are the predictions of the mixed relative-similarity/absolute-similarity exemplar model; see the Theoretical Analyses section.

the pattern of goodness-of-example judgments. To confirm these observations, statistical tests were conducted in which the results for individual subjects were used as the units of analysis. For each subject, the frequency with which stimulus $x\bar{y}$ (or $\bar{x}y$) was preferred to stimulus xy was computed. Because there were six paired comparisons involving $x\bar{y}$ (or $\bar{x}y$) versus xy , the null hypothesis of random responding leads to a mean expectation of three preferences for $x\bar{y}$ (or $\bar{x}y$). The observed means were 4.10, 2.37, and 1.43, in Conditions P_{00} , P_{10} , and P_{33} , respectively. The differences among the groups were statistically significant [$F(2,87) = 13.16$, $MS_e = 4.17$, $p < .01$]. Preference for $x\bar{y}$ (or $\bar{x}y$) was significantly greater than chance in Condition P_{00} [$t(29) = 2.82$, $p < .01$] and significantly less than chance in Condition P_{33} [$t(29) = -4.42$, $p < .001$].

The preference for stimuli $x\bar{y}$ and $\bar{x}y$ over stimulus xy in Condition P_{00} is inconsistent with independent feature-frequency models. More important to the theme of the present investigation is that the preference for $x\bar{y}$ (or $\bar{x}y$) is inconsistent with the joint hypothesis that subjects represented the positive concept solely in terms of the simple disjunctive rule and based goodness-of-example judgments on the degree to which the rule was instantiated. Conceivably, however, some subjects may have adopted unnecessarily complex rules for representing the positive concept (Medin, Wattenmaker, & Michalski, 1987; Nosofsky et al., 1989). For example, because stimulus xy was never presented in Condition P_{00} , some subjects may have adopted an exclusive-or rule. However, the exclusive-or rule was unavailable in Conditions P_{10} and P_{33} , because had subjects adopted it, they would have been unable to correctly classify xy during the final trials (20–32) of the criterion run (see the Method section). Even in the latter conditions taken by themselves, there was an effect of presentation frequency of stimulus xy on subjects' goodness-of-example judgments [$t(58) = 1.81$, $p < .05$, one-tailed test]. These effects of exemplar frequency on subjects' goodness-of-example judgments support the hypothesis that stored exemplars form an important component of the category representation, even for the present logically defined concept.

Table 5
Mean Typicality Ratings for Each Type of Stimulus as a
Function of Conditions in Experiment 2

Stimulus Type	Condition		
	P_{00}	P_{10}	P_{33}
xy	4.98	7.10	7.74
$x\bar{y}$ (or $\bar{x}y$)	7.22	6.98	6.15
$\bar{x}\bar{y}$	2.06	1.36	2.28

Typicality Ratings

The 16 typicality ratings obtained for each subject were averaged to form three ratings corresponding to xy , $x\bar{y}$ (or $\bar{x}y$), and $\bar{x}\bar{y}$ stimuli. The mean ratings for these three stimulus types are shown as a function of conditions in Table 5. The pattern of typicality ratings mirrors the paired-comparison judgments. In Condition P_{00} , ratings for the $x\bar{y}$ (or $\bar{x}y$) stimuli exceeded those for xy ; the ratings were roughly equal in Condition P_{10} ; and ratings for xy exceeded those for $x\bar{y}$ (or $\bar{x}y$) in Condition P_{33} . A 3×3 analysis of variance with conditions (P_{00} , P_{10} , P_{33}) and stimuli (xy , $x\bar{y}$, $\bar{x}\bar{y}$) as factors revealed that this interaction of typicality ratings with learning conditions was highly significant [$F(4,174) = 12.07$, $MS_e = 2.90$, $p < .01$]. This dramatic influence of exemplar frequency on the goodness-of-example judgments is consistent with a model in which it is assumed that individual exemplars form an important component of the mental representation for logically defined concepts.

Theoretical Analysis

The exemplar model was used to generate quantitative predictions of the paired-comparison judgments. The details of the model-fitting analyses are presented in Appendix C. The upshot of the analyses was that, as in Experiment 1, the mixed exemplar model involving both relative-similarity and absolute-similarity components yielded adequate predictions of the goodness judgments. The predictions of a three-parameter version of the model are shown along with the observed data in Table 4. Although the model is rejected in a statistical test of overall goodness-of-fit [$\chi^2(12) = 29.1$, $p < .05$], by conventional criteria it performs fairly well, accounting for 96% of the variance in the preference probabilities. Most notably, the mixed model accurately captured the crossover effect in rows 1 and 2 of the paired-comparison matrix, namely the preference reversal involving stimulus $x\bar{y}$ (or $\bar{x}y$) and stimulus xy as the presentation frequency of the latter was increased.

GENERAL DISCUSSION

The theme of this research involved the demonstration that even for logically defined concepts structured according to salient rules, stored exemplars may form a major component of the category representation. In previous work, Nosofsky (1984, 1986, 1989) showed that an exemplar-storage model incorporating selective attention processes

was capable of yielding precise quantitative predictions of classification performance for logically defined categories. The present work suggests that it is not simply a case of the selective-attention exemplar model "mimicking" rule-governed performance—people's mental representations of logically defined concepts may often be characterized better in terms of exemplar storage than in terms solely of storage of simple rules (see also Nosofsky et al., 1989). The main support for this proposal in the present research was that for both a unidimensional-rule category structure (Experiment 1) and a disjunctive-rule category structure (Experiment 2), people's goodness-of-example judgments were predicted better by principles of exemplar-based similarity than by the degree to which the logical rules were instantiated. Exemplar frequency was also demonstrated to affect speeded classifications using the unidimensional category structure tested in Experiment 1.

The claim in this article is not that people cannot apply and use simple rules when instructed to do so. Indeed, in explicit rule-instruction conditions, Nosofsky et al. (1989) observed patterns of classification performance that dramatically violated the predictions of the selective-attention exemplar model and that conformed to the predictions of rule-based models. What is claimed is that in learning logically defined concepts through induction from training exemplars, memories for the individual exemplars may form a dominant component of the category representation. The "stubbornness" of previously experienced exemplars can be viewed even in explicit rule-instruction conditions, in which application of the rule is often hampered (or facilitated) by similarities of probes to the specific exemplars on which training occurred (e.g., Brooks, 1987; Nosofsky et al., 1989, Experiment 1B; Ward & Scott, 1987).

As emphasized in the introduction, we have contrasted the exemplar model predictions with those of *particular* logical-rule models. These rule models seemed to be salient candidates; they were motivated by an extensive literature emphasizing the role of simplicity and cognitive economy in the formation of rules. Obviously, one approach to saving the rule-instantiation hypothesis in the present study would be to argue that people adopted logical rules, but not the ones assumed a priori by the experimenter. The danger with this approach, however, is that if the construct of a rule is left too vague, it fails to make predictions that can be falsified; that is, it is not testable. Indeed, virtually any pattern of classification performance and typicality judgments could be described in some post hoc manner by some form of verbalizable rule. With regard to the present experiments, we believe that the particular rules that were investigated were highly salient and plausible candidates. Furthermore, in informal interviews conducted after the experiment, essentially all subjects verbalized the small/large rule when describing the category structure that they learned in Experiment 1. Still, the possibility remains that more sophisticated rule-based models may be developed that would

predict that people would adopt more complex rules, and the rule-instantiation hypothesis may fare better under such circumstances (e.g., Medin et al., 1987).

The emphasis in this article has been on the exemplar side of category representation, but it would be implausible to posit that principles of rule instantiation play no role in the representation of logically defined concepts. In the domain of natural categories, for example, Barsalou (1985) has emphasized that many factors influence typicality judgments. In addition to factors involving exemplar frequencies and similarities, Barsalou (1985) demonstrated the importance of "ideals," defined as characteristics that items should have to satisfy goals associated with a category. As noted by Vandierendonck (1988, p. 79), the extent to which an item instantiates a logical rule seems similar in spirit to the construct of an ideal. Indeed, the rule-instantiation construct may be viewed as a form of "logical ideal," and it would not be surprising to see clear effects of ideals in logically defined concepts similar to the effects seen in natural categories.

Finally, the discussion of the present results would not be complete without reference to the well-known study of Armstrong, Gleitman, and Gleitman (1983), who demonstrated typicality gradients in well-defined semantic concepts such as "even numbers." One of Armstrong et al.'s main points was that the mere existence of typicality effects does not in and of itself establish that a concept is ill defined. But what of the psychological representation of the concept *even numbers*? The view suggested by the present work is that a highly derived, abstract psychological dimension, something like "divisible by 2," is selectively weighted when people make categorization decisions. But the vast milieu of people's specific experiences and interactions with even numbers is not thrown away. Highly frequent and salient exemplars such as 2, 4, and 10 are part and parcel of the category representation for even numbers, just as were the high-frequency exemplars for the logically defined concepts investigated herein.

REFERENCES

- ARMSTRONG, S. L., GLEITMAN, L. R., & GLEITMAN, H. (1983). What some concepts might not be. *Cognition*, **13**, 263-308.
- ASHBY, F. G., & GOTT, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.
- 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. (in press). A response time theory of perceptual independence. To appear in J. P. Doignon & J. C. Falmagne (Eds.), *Mathematical psychology: Current developments*. Berlin: Springer.
- ASHBY, F. G., & TOWNSEND, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154-179.
- BARSALOU, L. W. (1981). *Determinants of graded structure in categories*. Unpublished doctoral dissertation, Stanford University.
- BARSALOU, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 629-654.
- BOURNE, L. E., JR. (1982). Typicality effects in logically defined categories. *Memory & Cognition*, **10**, 3-9.
- BROOKS, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 141-174). New York: Cambridge University Press.
- CARROLL, J. D., & WISH, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 57-105). San Francisco: W. H. Freeman.
- CARTWRIGHT, D. (1941). Relation of decision-time to the categories of response. *American Journal of Psychology*, **54**, 174-196.
- EMMERICH, D. S., GRAY, J. L., WATSON, C. S., & TANIS, D. C. (1972). Response latency, confidence, and ROCs in auditory signal detection. *Perception & Psychophysics*, **11**, 65-72.
- ENNIS, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General*, **117**, 408-411.
- ENNIS, D. M., PALEN, J., & MULLEN, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology*, **32**, 449-465.
- ESTES, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.
- ESTES, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, **115**, 155-174.
- FESTINGER, L. (1943). Studies in decision: I. Decision time, relative frequency of judgment, and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, **32**, 291-306.
- GARNER, W. R. (1974). *The processing of information and structure*. New York: Wiley.
- GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- HAYES-ROTH, B., & HAYES-ROTH, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning & Verbal Behavior*, **16**, 321-328.
- HYMAN, R., & FROST, N. H. (1975). Gradients and schema in pattern recognition. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and Performance V* (pp. 630-654). London: Academic Press.
- KELLOGG, R. T. (1981). Feature frequency in concept learning: What is counted? *Memory & Cognition*, **9**, 157-163.
- LEEUEWENBERG, E. L. J. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, **84**, 307-349.
- LUCE, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.
- MEDIN, D. L., & SCHAFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
- MEDIN, D. L., WATTENMAKER, W. D., & MICHALSKI, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, **11**, 299-339.
- NEISSER, U., & WEENE, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology*, **64**, 640-645.
- NEUMANN, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, **2**, 241-248.
- NORMAN, D. A., & WICKELGREN, W. (1969). Strength theory and decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology*, **6**, 192-208.
- NOSOFSKY, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.
- NOSOFSKY, R. M. (1985a). Luce's choice model and Thurstone's categorical judgment model compared: Kornbrot's data revisited. *Perception & Psychophysics*, **37**, 89-91.
- NOSOFSKY, R. M. (1985b). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, **38**, 415-432.

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

NOSOFSKY, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 87-108.

NOSOFSKY, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.

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

NOSOFSKY, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, **45**, 279-290.

NOSOFSKY, R. M. (in press). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*.

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.

PIKE, A. R. (1973). Response latency models for signal detection. *Psychological Review*, **80**, 53-68.

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

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

REITMAN, J. S., & BOWER, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, **4**, 194-207.

RESTLE, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review*, **86**, 1-24.

SHEPARD, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325-345.

SHEPARD, R. N. (1958a). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, **65**, 242-256.

SHEPARD, R. N. (1958b). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, **55**, 509-523.

SHEPARD, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**, 54-87.

SHEPARD, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General*, **115**, 58-61.

SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

SHEPARD, R. N., HOVLAND, C. I., & JENKINS, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**(13, Whole No. 517).

SMITH, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, **69**, 77-110.

SUPPES, P. (1957). *Introduction to logic*. Princeton, NJ: Van Nostrand.

THOMAS, E. A. C. (1971). Sufficient conditions for monotone hazard rate: An application to latency-probability curves. *Journal of Mathematical Psychology*, **8**, 303-332.

THURSTONE, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273-286.

VANDIERENDONCK, A. (1988). Typicality gradient in well-defined artificial categories. *Acta Psychologica*, **69**, 61-81.

WARD, T. B., & SCOTT, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, **15**, 42-54.

WICKELGREN, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, **5**, 102-122.

NOTES

1. Unexplainedly, only a small effect was observed in Condition L. However, in a direct replication involving the same numbers of sub-

jects, significant effects of the angle manipulation were observed in Condition L. Thus, the null result in Condition L of the present experiment is probably best attributed to sampling error.

2. As shown in Figure 7, the absolute magnitudes of the RTs for the Size 4 and Size 5 stimuli were greater in Condition HF12/L than in Condition HF12/L. This result is part of a larger pattern in which RTs for the Category B stimuli tend to be longer in Condition HF12/L than in Condition HF12/L, whereas RTs for the Category A stimuli tend to be shorter (see Table 2). This pattern may be the result of a response-bias shift across conditions. With the addition of Stimuli 15 and 16 to the training set, the proportion of Category B presentations during training increased slightly in the latter condition and may have induced more of a readiness to respond with Category B during the speeded classifications. The important result concerns the relative magnitude of RTs for the Size 4 and Size 5 stimuli across conditions, not their absolute magnitudes.

APPENDIX A

Typicality Paired-Comparison Data, Experiment 1A

A. Condition HF10/L Matrix								
	7	9	10	12	13	14	15	
7		28	5	10	22	25		
9	22		9	1	22	24		
10	45	41		29	45	46	30	
12	40	49	21		42	50		
13	28	28	5	8		32		
14	25	26	4	0	18			
15			20					
B. Condition HF12/L Matrix								
	7	9	10	12	13	14	16	
7		26	6	5	30	28		
9	24		11	7	33	29		
10	44	39		24	45	41		
12	45	43	26		46	43	36	
13	20	17	5	4		25		
14	22	21	9	7	25			
16				14				
C. Condition HF10/L Matrix								
	7	9	10	12	13	14	15	
7		29	6	11	32	31		
9	21		10	3	30	25		
10	44	40		30	47	41	13	
12	39	47	20		45	47		
13	18	20	3	5		29		
14	19	25	9	3	21			
15			37					
D. Condition HF12/L Matrix								
	7	9	10	12	13	14	16	
7		19	8	13	39	39		
9	31		17	8	35	36		
10	42	33		14	45	44		
12	37	42	36		47	45	13	
13	11	15	5	3		26		
14	15	14	6	5	24			
16				37				
E. Condition HF15 Matrix								
	7	9	10	12	13	14	15	16
7		29	7	27	40	39		
9	21		14	8	35	36		
10	43	36		33	47	48	1	
12	23	42	17		44	47		

	7	9	10	12	13	14	15	16
13	10	15	3	6		26		
14	11	14	2	3	24			
15			49					29
16							21	

F. Condition HF16 Matrix

	7	9	10	12	13	14	15	16
7		21	8	11	38	30		
9	29		21	11	35	30		
10	42	29		19	50	40		
12	39	39	31		44	44		3
13	12	15	0	6		28		
14	20	20	10	6	22			
15								12
16				47			38	

Note—Entry in each cell gives the frequency with which the row stimulus was judged as a better example of Category B than was the column stimulus in each condition. Entries in cell (i, j) and cell (j, i) sum to 50.

APPENDIX B

Experiment 1A Theoretical Analyses

In this appendix, the exemplar model is tested on its ability to make accurate quantitative predictions of the typicality ratings and paired-comparison data in Experiment 1A. To apply the model, a psychological scaling solution for the exemplars is needed (e.g., see Nosofsky, 1986, 1987). The scaling solution is derived by fitting the *MDS-choice* model to the set of confusion data obtained in the identification study (Nosofsky, 1985a, 1985b, 1987, 1989; Shepard, 1958b).

MDS-Choice Model Analysis

The data obtained during Trials 21–150 of the identification experiment are summarized in the confusion matrices shown in Table B1. For simplicity in analyzing and reporting the results, and because the complete 25×25 matrix of identification confusion data was sparse, the data were collapsed into two 5×5 submatrices corresponding to size and angle. For example, cell (2,3) of the size matrix gives the frequency with which Size 2 stimuli were identified as Size 3 stimuli. In general, the data in each matrix show a regular pattern in which confusability decreases as separation between levels on each dimension increases (e.g., values 2 and 4 show less confusions than values 2 and 3). This pattern supports the assumption that the physical variations along each dimension map onto essentially unidimensional psychological continua. An exception occurs, however, for Angles 1 and 5, which actually show greater confusability than, for example, Angles 1 and 4, and so forth. This pattern indicates that the appropriate psychological representation for angle is at least two-dimensional, with the second dimension corresponding, perhaps, to angle extremity. For simplicity in the theoretical analyses, however, the MDS-choice model was fitted to the confusion data on the assumption that there were no confusions between Angles 1 and 5. Although incorrect, this assumption greatly simplifies the presentation of the ensuing theoretical analyses and leaves unchanged all major conclusions regarding prediction of the typicality data.

The MDS-choice model was fitted separately to each confusion matrix by searching for the unidimensional scaling solution that maximized the likelihood of the data. Specifically, for the angle matrix, each level of angle (j) was represented by a psychological coordinate, a_j ; the distance between angle levels i and j was given by $d_{ij} = |a_i - a_j|$; and these distances were

then transformed to similarity measures, using Equation 2. Two values of p in the similarity function were considered: $p = 1$, which yields an exponential decay function; and $p = 2$, which yields a Gaussian function. The derived similarity values (s_{ij}) were then substituted into the choice model equation (Luce, 1963; Shepard, 1957) to predict the probability that angle i was identified as angle j :

$$P(R_j | S_i) = \frac{b_j s_{ij}}{\sum_k b_k s_{ik}}, \quad (\text{B1})$$

where b_j is a parameter reflecting the bias for angle-response j . An analogous procedure was used for fitting the MDS-choice model to the size matrix.

The maximum-likelihood predicted confusion frequencies for the exponential-similarity choice model are shown alongside the observed frequencies in Table B1. As can be seen, the fits are good. The maximum-likelihood parameters and summary fits are reported in Table B2. For purposes of comparison, the summary fits for a version of the MDS-choice model in which a Gaussian similarity function was assumed are also shown in Table B2. The Gaussian model has been strongly favored in previous work involving the identification of highly confusable stimuli (e.g., Nosofsky, 1985a, 1985b, 1989), but the exponential model is clearly favored in the present study, in which fairly discriminable stimuli were used. This result lends further support to proposals put forth by Shepard (1986, 1987) and Ennis (1988; Ennis et al., 1988) regarding the role of stimulus discriminability in influencing the form of the similarity gradient.

A two-dimensional scaling solution for the 25 stimuli, formed by combining orthogonally the derived unidimensional scaling solutions for angle and size, is illustrated in Figure 2. This two-

Table B1
Angle and Size Identification Confusion Matrices

Stimulus	Response				
	1	2	3	4	5
Angle Matrix					
1	608 (608)	8 (8)	3 (4)	1 (0)	23 (0)
2	0 (0)	467 (467)	231 (224)	22 (29)	1 (1)
3	1 (0)	47 (55)	631 (630)	86 (83)	4 (1)
4	2 (0)	14 (11)	118 (125)	532 (527)	4 (7)
5	19 (0)	5 (0)	3 (4)	15 (18)	665 (665)
Size Matrix					
1	609 (609)	94 (85)	8 (15)	1 (3)	0 (0)
2	107 (99)	480 (497)	109 (87)	1 (18)	4 (0)
3	9 (15)	85 (73)	467 (488)	114 (99)	2 (2)
4	6 (8)	39 (41)	286 (274)	382 (385)	2 (6)
5	0 (0)	1 (2)	9 (15)	28 (21)	667 (667)

Note—Values in parentheses are the maximum-likelihood predicted frequencies for the exponential-similarity MDS-choice model. The model was fitted to the data on the assumption that there were zero confusions between Angles 1 and 5.

Table B2
Maximum-Likelihood Parameters and Summary Fits
for the MDS-Choice Model

	Exponential Model		Gaussian Model	
	Angle Matrix			
j	a_j	b_j	a_j	b_j
1	0	.010	0	.007
2	7.137	.156	2.742	.166
3	8.725	.366	3.869	.349
4	10.461	.273	5.059	.301
5	14.389	.195	7.141	.176
Fits:	$-\ln L = 55.0$ $SSE = 318.73$		$-\ln L = 150.0$ $SSE = 8,483.33$	
	Size Matrix			
	s_j	b_j	s_j	b_j
1	0	.280	0	.306
2	1.792	.235	1.237	.248
3	3.611	.253	2.356	.236
4	4.577	.135	3.128	.135
5	8.388	.097	5.000	.074
Fits:	$-\ln L = 76.3$ $SSE = 2,355.82$		$-\ln L = 134.0$ $SSE = 10,191.98$	

Note—ln *L* = log-likelihood, SSE = sum of squared deviations between predicted and observed confusion frequencies. (Log-likelihood was the criterion of fit.)

dimensional scaling solution was used in conjunction with the exemplar-similarity model to predict the typicality data in the classification learning experiment.

Exemplar-Model Predictions of the Typicality Data

As currently developed, the exemplar model is directly applicable to the prediction of classification probabilities, but the present data involve ratings and paired comparisons. Extensive exploratory analyses were conducted to find transformations of underlying variables in the exemplar model that accurately predicted the typicality ratings and paired comparisons. The upshot of these analyses was that a mixture of the relative- and absolute-similarity measures was necessary to adequately predict the data.

Typicality ratings. The two-dimensional scaling solution for the exemplars that was derived from the identification confusion data was used in conjunction with the exemplar model to predict the typicality ratings. Because the stimuli varied along readily discriminable, separable dimensions (Garner, 1974; Shepard, 1964, 1987), the distance between exemplars *i* and *b* was computed using a weighted city-block metric:

$$d_{ib} = c[w_1|x_{i1} - x_{b1}| + (1 - w_1)|x_{i2} - x_{b2}|], \quad (B2)$$

where x_{ik} is the psychological value of exemplar *i* on dimension *k*. (These psychological values correspond to the derived angle and size coordinates reported in Table B2.) In Equation B2, *c* is a general sensitivity parameter, and w_1 ($0 \leq w_1 \leq 1$) is the weight given to Dimension 1 (angle) in computing distance. The distance was transformed to a similarity measure (s_{ib}), using the same exponential similarity function that was used in the scaling. These derived similarity measures were then substituted into Equation 1 to predict the probability with which exemplar *i* was classified into Category B. In deriving the predictions, the exemplar-strength terms (N_b in Equation 1)

were set at: 0, if exemplar *b* had not been presented during training; 1, if exemplar *b* had been presented with normal frequency during training; a free parameter, *h*, if exemplar *b* had been presented with high frequency during training. Finally, the predicted Category B response probability was transformed to a standard normal deviate, $z[P(R_B|S_i)]$. The reason for using this transformation is that the probability measure has a floor and ceiling (at 0.0 and 1.0), and the *z* transformation effectively stretches the scale at these regions.

As discussed previously, the classification probability, $P(R_B|S_i)$, is based on a relative-similarity rule, whereas typicality judgments may often involve an absolute summed-similarity component. The absolute summed similarity of exemplar *i* to the exemplars of Category B was given by

$$A(i, B) = \sum_{b \in B} N_b s_{ib}, \quad (B3)$$

where N_b and s_{ib} were computed as before.

The *typicality strength* for exemplar *i*, $T(i)$, was assumed to be based on a mixture of the relative-similarity and absolute-similarity components:

$$T(i) = m \cdot z^*[P(R_B|S_i)] + (1 - m) \cdot A^*(i, B), \quad (B4)$$

where *m* is a freely estimated mixture parameter. The asterisks in Equation B4 denote that both the relative and absolute-similarity components were standardized to have a mean of 0 and standard deviation of 1 in each individual condition. It was assumed that the typicality ratings would be monotonically related to the typicality strengths defined by Equation B4.

In summary, the typicality-strength model has four free parameters: the sensitivity parameter *c* and attention weight *w* in the distance function; the exemplar-strength *h* associated with high-frequency exemplars; and the mixture parameter *m*. A computer search was conducted to find the values of these four parameters that maximized the Spearman rank-order correlation between typicality strengths and observed typicality ratings for all 16 stimuli across all six conditions. (All parameters were held fixed across conditions.) With $c = .7$, $w = .2$, $h = 1.3$, and $m = .7$, the model achieved a rank-order correlation of .974.

For purposes of comparison, special cases of the mixed model were fitted to the data. A pure relative-similarity model ($m = 1$ in Equation B4) yielded a rank-order correlation of $r = .84$. This model incorrectly predicted high typicality ratings for the Size 5 stimuli in Conditions HF10/L and HF12/L and for Stimuli 13 and 14 across all six conditions. A pure absolute-similarity model ($m = 0$ in Equation B4) yielded a rank-order correlation of .91. This model incorrectly predicted low typicality ratings for the Size 5 stimuli in Conditions HF10/L and HF12/L, and also incorrectly predicted that in Conditions HF15 and HF16, typicality ratings for the Size 3 and Size 4 stimuli would be virtually identical.

Typicality Paired Comparisons

The typicality-strength model was also used to predict the paired-comparison typicality judgments. For each condition, the probability that exemplar *i* was judged as a better member of Category B than was exemplar *j*, $P(i, j)$ was given by the logistic transformation

$$P(i, j) = \{1 + \exp[v \cdot (T(j) - T(i))]\}^{-1}, \quad (B5)$$

where *v* is a freely estimated scaling parameter. This five-parameter model (*c*, *w*, *h*, *m*, *v*) was fitted to the data in the six paired-comparison matrices, using a minimum chi-square

criterion. Note that, across all six conditions, there were 98 paired comparisons. The best-fitting parameters ($c = .96$, $w = .74$, $h = 1.17$, $m = .84$, $v = 4.08$) yielded $\chi^2(93, N = 4,900) = 217.5$, which is significant at the $p = .001$ level. (Because the paired-comparison data contain multiple observations from individual subjects, the independence assumption underlying the chi-square test is undoubtedly violated. Thus, the results of the statistical tests of the model should be interpreted with caution.) The special-case models fared dramatically worse. The pure relative-similarity model yielded a chi-square of 1,444.1; and the pure absolute-similarity model, a chi-square of 631.5.

A quantitative version of the rule-based model was also fitted to the paired-comparison data. The rule model assumes that goodness-of-example judgments are determined by the degree to which the rule is instantiated—that is, by the distance of an item from a boundary that separates small from large stimuli. Let $S(i)$ denote the psychological value of size associated with stimulus i (given by the coordinates in Table B2), and let B be a freely estimated parameter corresponding to the value of the boundary. According to the rule model, the typicality strength associated with stimulus i is given by

$$T(i) = S(i) - B. \quad (B6)$$

Application of the logistic transformation (Equation B5) then yields the predicted paired-comparison probabilities. This two-parameter model failed dramatically at predicting the paired-comparison judgments, yielding a chi-square of 1,517.1. Among the rule model's major shortcomings were that it predicted high typicality strength for Stimuli 13 and 14 across all conditions and high typicality strength for the Size 5 stimuli in Condition $\bar{L}$, and that it was insensitive to the effects of the frequency manipulation on the irrelevant angle dimension.

It was unclear how to add theoretically motivated free parameters to the rule model in order to achieve fairer comparisons with the exemplar model. The main conclusion is simply that the present quantitative model-fitting analyses corroborate the previously reported qualitative analyses in suggesting that subjects' mental representations did not consist solely of an abstract rule boundary for partitioning the members of the alternative categories.

APPENDIX C

Experiment 2 Theoretical Analyses

In this appendix, the application of the exemplar model to the typicality paired-comparison data obtained in Experiment 2 is

outlined. A similarity parameter s was defined that corresponded to each of the two relevant dimensions, and, following Medin and Schaffer (1978), an interdimensional multiplicative rule was used to compute the overall similarity between exemplars. Specifically, the similarity between exemplars i and j was given by

$$s_{ij} = \prod_{m=1}^2 s^{\delta_m(i,j)}, \quad (C1)$$

where $\delta_m(i,j)$ is an indicator variable equal to one if exemplars i and j mismatch on dimension m , and equal to zero otherwise. (This multiplicative rule is a special case of the exponential/city-block model of similarity used to analyze the Experiment 1A data—see Nosofsky, 1984, 1986.) So, for example, the similarity between stimuli xy and $\bar{x}\bar{y}$ equals s ; the similarity between xy and $\bar{x}y$ equals s^2 ; and the similarity between any stimulus and itself equals one. These derived similarities were then substituted into Equation 1 to predict the probability with which an exemplar will be classified into the positive concept. For simplicity, the exemplar-strength terms in Equation 1 were assumed to be proportional to the relative frequency with which the exemplars were presented during training. So, for example, in Condition P_{10} , the probability with which xy is predicted to be classified as a member of the positive concept would be given by

$$P(R+ | xy) = \frac{.05 + .45s}{(.05 + .45s) + .5s^2}. \quad (C2)$$

As was the case in Experiment 1A, these predicted probabilities were transformed to standard normal deviates, $z[P(R+ | S_i)]$. The absolute summed similarity of an item to the exemplars of the positive concept (i.e., the numerator in the classification decision rule) was also computed. As in Experiment 1A, the typicality strength for an item was given by the mixture model (Equation B4), and the paired-comparison choice probabilities were given by the logistic transformation (Equation B5). The model required estimation of three parameters—the similarity parameter s in Equation C1, the mixture parameter m in Equation B4, and the logistic scaling parameter v in Equation B5—to predict the 15 paired-comparison probabilities in Table 3. The minimum chi-square parameters were $s = .30$, $m = .12$, and $v = 9.35$, yielding $\chi^2(12, N=1,350) = 29.1$.

(Manuscript received May 15, 1989;
revision accepted for publication October 4, 1990.)