Journal of Experimental Psychology: General
1980, Vol. 109, No. 3, 320-353

Classification in Well-Defined and Ill-Defined Categories: Evidence for Common Processing Strategies

Randi C. Martin and Alfonso Caramazza
Johns Hopkins University

SUMMARY

Early work in perceptual and conceptual categorization assumed that categories had criterial features and that category membership could be determined by logical rules for the combination of features. More recent theories have assumed that categories have an ill-defined structure and have proposed probabilistic or global similarity models for the verification of category membership.

In the experiments reported here, several models of categorization were compared, using one set of categories having criterial features and another set having an ill-defined structure. Schematic faces were used as exemplars in both cases. Because many models depend on distance in a multidimensional space for their predictions, in Experiment 1 a multidimensional scaling study was performed using the faces of both sets as stimuli.

In Experiment 2, subjects learned the category membership of faces for the categories having criterial features. After learning, reaction times for category verification and typicality judgments were obtained. Subjects also judged the similarity of pairs of faces. Since these categories had characteristic as well as defining features, it was possible to test the predictions of the feature comparison model (Smith et al.), which asserts that reaction times and typicalities are affected by characteristic features. Only weak support for this model was obtained. Instead, it appeared that subjects developed logical rules for the classification of faces. A characteristic feature affected reaction times only when it was part of the rule system devised by the subject.

The procedure for Experiment 3 was like that for Experiment 2, but with ill-defined rather than well-defined categories. The obtained reaction times had high correlations with some of the models for ill-defined categories. However, subjects' performance could best be described as one of feature testing based on a logical rule system for classification.

These experiments indicate that whether or not categories have criterial features, subjects attempt to develop a set of feature tests that allow for exemplar classification. Previous evidence supporting probabilistic or similarity models may be interpreted as resulting from subjects' use of the most efficient rules for classification and the averaging of responses for subjects using different sets of rules.

Human beings make sense of variability in the perceptual world by classifying similar objects into the same category. Psychological research has long been directed toward determining the structure of the similarity relations among members of a class or category and the process by which exemplars are classified. Different class or category types ranging in complexity from simple geometric forms (triangle, circle) to superordinate lexical categories (furniture, clothing) have been used. The assumption has often been made either explicitly or implicitly that the same structural and processing principles hold in both perceptual and conceptual realms and that, therefore, one can learn about conceptual categorization by studying in the laboratory the classification of simple perceptual forms (Bruner, 1957; Rosch & Mervis, 1975). Consequently,

parallel theoretical developments have taken place in research on perceptual and conceptual categorization.

In recent theoretical and empirical developments it has been assumed that most natural concepts are ill-defined—that is, that there is no simple set of criterial features that can be used to determine membership of all exemplars of a category. Given this assumption, efforts have focused on developing probabilistic models of categorization or models that depend on global similarity between an exemplar and a category representation to determine category membership. Cue validity, average distance, and prototype models have been proposed that do not depend on the strong assumption that there is a specific set of meaning components or features that are both necessary and sufficient for classification of exemplars. Interestingly, however, whatever evidence has been presented against the class of models based on criterial features testing has been indirect and, as will be seen, does not constitute sufficient grounds against this class of models. In this article we report a series of experiments that compare a sequential feature testing model to the probabilistic and global similarity models. Specifically, we attempt to show that if individual differences are taken into consideration, a sequential feature testing model provides a powerful basis for a processing model of classification.

Perceptual Categories

Early work in the area of classifying perceptual forms went under the label *concept formation* (Bruner, Goodnow, & Austin, 1956; Neisser & Weene, 1962). This work made two major assumptions: (a) that the objects to be classified could be analyzed into well-defined features and (b) that the basis for classification was a rule for combination of these features (Bourne, 1970). Subjects in these studies were presented with stimuli varying along obvious dimensions (such as size or color), and their task was to learn the rule that determined category membership for each object (for example, all objects in Category 1 were small and red, whereas those in Category 2 were large and blue). A typical issue in these studies was to determine how subjects went about finding these regularities. Although differences in strategy were observed, the process by which subjects learned the rules has been accepted to be one of hypothesis testing (Bourne, Ekstrand, & Dominowski, 1971). Based on a positive instance, the subject makes a guess as to what the rule is and uses the hypothesized rule in making subsequent responses until she or he encounters an instance that disconfirms the rule.

Later researchers questioned either or both of the major assumptions of the concept formation studies. Some claimed that perceptual objects in the real world could not be broken down into well-defined features and, consequently, that there were no logical rules for feature combination that determined an object's classification (Posner, Goldsmith, & Welton, 1967; Posner & Keele, 1968). Others did not object so much to the idea of objects' being composed of separable features but did object to the assumption that there was some simple logical combination of features that would determine category membership for all exemplars (Reed, 1972).

In the studies conducted by Posner and his colleagues, stimuli lacking both well-defined features and logical rules for classification were used. The categories used were constructed so that exemplars were distortions of a random configuration of dots that served as the prototype configuration for a category. Posner and his colleagues were interested in determining how subjects learned the categories and how they classified new exemplars. Posner and Keele (1968) found that in transfer tasks subjects could classify new distortions better if their original learning had been on more, as opposed to less, distorted pat-

The research reported here was supported in part by National Institute of Health Research Grant 14099. We wish to thank Michael McCloskey and Howard Egeth for their comments on an earlier draft of this article. We also thank Michael Giordano and Kevin Stone for their assistance in collecting the data.

Requests for reprints should be sent to Randi C. Martin, Department of Psychology, Johns Hopkins University, Baltimore, Maryland 21218.

terns. Also, they found that the prototype of the category (not presented during the learning phase) could be classified as well as the exemplars that had been presented during the learning phase. From these results Posner and Keele concluded that during the learning phase subjects abstracted the prototype for the category from the distortions but also retained information about the expected variability of the exemplars from this prototype.

In contrast to the Posner studies, Reed (1972) used stimuli that had well-defined features (faces with nose, mouth, etc.) that varied continuously (length of nose, distance between eyes, etc.). To obtain a representation of the psychological similarity of the stimuli, Reed performed a multidimensional scaling analysis of similarity judgments of stimulus pairs. Members of a category were selected randomly except for the constraint that the categories could be separated by a linear discriminant function in the multidimensional space derived for the stimuli. Given that there was no simple, logical rule for classifying stimuli (for example, no single feature value was common to all category members), Reed was interested in determining what information about category members subjects would use in classifying new exemplars. He considered two general models, a probability model and a distance model. According to the probability model, subjects assign a stimulus to a category on the basis of the computed cue validity of each feature value, that is, the conditional probability that a face is in a category, given that it has a particular feature value. Reed proposed several variations of a cue validity model, such as models with differentially weighted features or models incorporating only the most predictive cues, but all depend on the frequency of particular values of features for determining category membership. The distance models he tested all depend on distance between stimuli in a multidimensional space for determining category membership, with stimuli being classified on the basis of their being near the average of a category (the prototype model) or being, on the average, closer to the members of one category than to the members of the other category (the average distance model). In all but one case, Reed found that subjects' classifications could best be predicted on the basis of one version of the distance model, specifically, the prototype model. The one exception occurred when the prototypes for the different categories were quite similar, in which case a cue validity model based on one cue was most predictive.

Thus, both Posner and Keele (1968) and Reed (1972) found that when subjects are presented with stimuli that do not allow some obvious rule for classification, subjects will tend to abstract a prototype for a category and use distance from it to predict category membership. However, as Reed found, some stimulus conditions do not lead to prototype abstraction. More recent work (Barresi, Robbins, & Shain, 1975; Goldman & Homa, 1977; Homa & Chambliss, 1975; Neumann, 1977) has identified the conditions of a stimulus set, such as degree of category overlap, variability within category, and degree of feature continuity, that lead to cue validity models' (feature counting models) making better predictions than prototype models (feature averaging models).

Whereas Posner and Keele (1968) and Reed (1972) were concerned with the basis of classifications of new exemplars, a study by Hyman and Frost (1975) attempted to discriminate among various models by studying reaction time for verification of category membership. Distortions of a single dot pattern were used as category exemplars. Distortions with greater height than width were assigned to one category and those with greater width than height to the other. After subjects had learned the category membership of the exemplars, reaction times for categorization were recorded. Two of the four competing models tested were exemplar models, that is, models that assume that the subjects use information about specific members of a category rather than general category information in determining an instance's category membership. The average exemplar model assumed that reaction time is a function of the difference between the average similarity of the instance to the

nearer category minus the average similarity of the instance to the farther category. The nearest exemplar model assumes that subjects find the one exemplar most similar to the present instance and classify the instance on the basis of the category membership of this nearest exemplar. Reaction time is assumed to be a function of the similarity of the nearest exemplar to the instance minus the similarity of the nearest exemplar in the contrasting category.

The third model, a prototype model, is very similar to that of Reed (1972). However, Hyman and Frost predicted that not only would similarity of the instance to the prototype of the nearer category affect reaction time but that the similarity to the prototype of the farther category would also have an effect. Specifically, reaction time was predicted to be a function of distance of the instance to the nearer prototype minus distance to the farther prototype. The fourth model assumed that subjects would abstract the rule, either consciously or unconsciously, that was used to structure the categories. Reaction time was predicted to be a function of the absolute value of the difference between height and width.

Hyman and Frost (1975) obtained high correlations between predicted and obtained reaction times for all models (ranging from $r = .72$ to $r = .87$ in one experiment). However, a series of converging operations indicated that for one of the pairs of contrasting categories they used (Design 1), the rule model was the best predictor of performance, whereas for a second set (Design 2), the prototype model made the best predictions.

Recently, Medin and Schaffer (1978) have proposed a context model that, like the average distance model of Reed and the average exemplar model of Hyman and Frost, assumes that categorization depends on stored information about category members rather than on overall category information. They assume that instances are categorized on the basis of their similarity to the exemplars of the competing categories. However, in contrast to other exemplar models, similarity is not determined by a sum of similarities for each feature but by a product. This implies that a specific exemplar will be classified more easily if it has high similarity to some exemplars and low similarity to others in the set than if it has medium similarity to all exemplars. Hence, the lack of success of other exemplar models does not preclude the possibility that the context model of Medin and Schaffer can account for results obtained in earlier studies.

Medin and Schaffer (1978) considered reaction time for both the classification of old exemplars and the classification of new exemplars in their experiments. They obtained results generally supportive of their model when testing their model against prototype or cue validity models that used a sum for computing similarity rather than a product of feature overlap. However, in their experiments they used very small categories, ranging in size from three to five exemplars. Because it seems less plausible that one would compare a new instance to all exemplars of possible categories when categories are large, Medin and Schaffer proposed that for large categories an instance brings to mind the exemplar to which it is most similar, and categorization is based on the membership of this nearest exemplar. No empirical evidence was presented to support this claim.

Present research on perceptual categorization is equivocal with regard to the first assumption of the early concept formation studies, that is, whether or not category exemplars can be decomposed into well-defined features. Both stimuli that are easily decomposable (e.g., schematic faces) and those that are not (e.g., dot configurations) are being used as exemplars. However, the second assumption, that category membership is rule determined, appears to have been abandoned. A typical concern of present-day research is to determine which of several competing models best describes subjects' behavior when they are confronted with categories for which there are no logical rules defining category membership.

Lexical Categories

The development of psychological research on conceptual categories has followed a pattern similar to that for per-

ceptual categories. Following the work of such semanticists as Katz and Fodor (1963) and Bierwisch (1970), psychologists adopted a componential view of meaning that proposes that a word can be broken down into elementary components, or semantic features, and that these components specify the necessary and sufficient features of the referents of the words. (The semanticists intended their componential theories to account for the meanings of all words, but for the present discussion we will be concerned only with the meanings of words that could be considered categories, that is, labels of concrete, physical objects.) Psychological research was directed toward empirical determination of the semantic features of domains of words, for example, animal terms (Henley, 1969), kinship terms (Romney & D'Andrade, 1964), and many others (Fillenbaum & Rapoport, 1971; Romney, Shepard, & Nerlove, 1972).

Objections to the componential formulation have been made (as with concept formation) on the grounds that the features of objects cannot be specified precisely and are not well-defined for most categories (Bolinger, 1965; Hutchinson & Lockhead, 1977) or that even though features may exist, there is no set of necessary features that will determine category membership for all category members (Rosch, 1973; 1975; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976; Wittgenstein, 1953).

Although the existence of semantic features has been questioned, there has also been a good deal of evidence indicating the theoretical usefulness of semantic features in explaining such semantic facts as anomaly and contradiction (Katz, 1972; Leech, 1974) and in explaining the acquisition of word meaning (Clark, 1973). Some researchers have preferred to speak in terms of operations rather than features (Miller & Johnson-Laird, 1976), but few have assumed that the meaning of a word cannot be broken down into components. Much more evidence indicates that there is no set of critical features that determines an object's membership in a category. An important finding in this regard is that not all members of a category are equally good members; some members of a category will be reliably judged to be more typical of a category than others. Degree of typicality has been shown to affect reaction time for verification of category membership, with more typical items being verified more quickly (Caramazza, Hersh, & Torgerson, 1976; Rips, Shoben, & Smith, 1973; Rosch & Mervis, 1975). If all category members had the same criterial features, one would expect all category members to be equally good members, since all would have the features necessary for category membership. Smith, Shoben, and Rips (1974) have circumvented this problem by assuming that words have not only criterial features (which they refer to as defining features) but also characteristic features, that is, features that are commonly possessed by category members but are not necessary for category membership. Members that have many of the characteristic features of a category are considered to be more typical of a category than those that do not.

Other researchers, such as Rosch (1973, 1975) and her colleagues (Rosch & Mervis, 1975; Rosch et al., 1976) and McCloskey and Glucksberg (1979) have done away entirely with the notion of defining features. They assume, as did Wittgenstein (1953), that category members bear a "family resemblance" to one another. Members of a language category are assumed to have similar features, but with some members of the category overlapping on some features and others overlapping on other features. For the category as a whole, some features will be more representative of the category than others, but no feature that can serve to distinguish category members from nonmembers will be common to all category members. Within this framework, the typicality of an exemplar is determined by the number of characteristic features it possesses.

Evidence for this position comes from a study by Rosch and Mervis (1975) in which subjects were asked to list the attributes of members of common categories such as clothing and furniture. Typicality judgments for the members of these categories were also obtained. Rosch and Mervis found that for four of the six categories used, only one listed attribute was common to all category

members, and for the remaining two categories, there were no common attributes listed. Moreover, for three of the four categories that did have a common attribute, the common attribute was not sufficient for categorizing exemplars into that category. Rosch and Mervis computed a family resemblance score for each member of each category by counting the number of times an attribute listed for an item was also listed for another item in the category and summed these counts over all attributes listed for that item. Typicality of the exemplar was found to be highly correlated with its family resemblance score.

Rosch concludes from these findings that categories do not have defining features but are instead organized in terms of family resemblance, with typicality of a category member determined by its feature overlap with other members of the category. However, it could be objected that asking subjects to list attributes biases them toward listing characteristic features of the category members, that is, salient perceptual attributes such as color, form, and size that vary from one member to the next. If subjects were indeed listing mostly characteristic features, then the correlations of typicalities and computed family resemblance scores could also be predicted on the basis of the Smith et al. model.

As in the case of perceptual categories, research with lexical categories has been concerned not only with the structure of categories but also with the process by which people categorize. Smith et al. (1974) have proposed a model that uses both defining and characteristic features in the process of verifying category membership. According to their model, in the first stage of processing, all the features of the exemplar are compared to all the features of the category (both defining and characteristic) in a rapid, all-at-once fashion. If there is a high degree of overlap or a very low degree of overlap, the second stage is not entered, and the decision for or against category membership can be made quickly. However, if there is a moderate degree of overlap, the second stage is employed, in which the features of the exemplar are checked against the defining features of the category. Thus, for the less typical category members, the second stage will be necessary and reaction time for verification will be longer. The fact that a continuous range of reaction times are obtained rather than two discrete levels (second stage vs. no second stage) is accounted for by assuming that the probability of entering the second stage varies in a continuous fashion dependent on the degree of overlap obtained in the first stage.

Within the family resemblance position, processing models similar to those proposed by Reed (1972), that is, cue validity and prototype models, have been entertained. However, unlike those in Reed's study, models developed for lexical categories are concerned with explaining reaction time for category verification rather than explaining which category an exemplar is placed in, since the lexical category membership of exemplars is assumed already to be known by the subject. Rosch and Mervis (1975) have, for the case of discrete features, proposed a model similar to a cue validity model, which assumes that it is an exemplar's overlap with the features of remaining category members that determines reaction time. The higher the degree of overlap, the faster the exemplar will be categorized. Rosch has presented no specific processing assumptions about how degree of overlap is computed by subjects, but one could assume either that the comparison is made on an instance by instance basis or that the subject has knowledge of the category structure (i.e., that three members have feature *a*, four have feature *b*, and so on) and that the comparison is made between an instance's features and these overall totals. For the case of continuous features, that is, when the values of features can be ordered along some continuum, Rosch et al. (1976) have proposed a prototype model in which it is distance from the prototype that determines reaction time for classification, with smaller distances being correlated with faster reaction times.

Assuming that the same kind of processing occurs with perceptual and conceptual categories, Rosch has obtained evidence for her claims through experiments using artificial categories composed of letter

strings and stick figures (Rosch et al., 1976). The letter strings were used as examples of category members with discrete features. A family resemblance score for each letter string was computed by summing for each letter in a string the number of other strings in the category having the same letter. In an experiment in which subjects were required to learn which category the letter strings belonged to, family resemblance was found to have a significant effect on rate of learning, reaction time for categorization, and typicality judgments made by the subjects after learning the categories. In the experiment in which the exemplars were stick figures with continuously varying features, Rosch et al. found that distance from the prototype of the category (a figure having the average value for the category for each feature) had a significant effect on rate of learning, reaction time, and typicality judgments.

Although these results are consistent with Rosch's proposals, they do not eliminate other possible models from contention. She has not demonstrated that one could not obtain these results if the categories had defining features but still differed in degree of feature overlap. Moreover, even assuming that categories do have a family resemblance structure, many other models, such as those of Medin and Schaffer and Hyman and Frost, would make predictions highly correlated with those based on family resemblance scores and hence would predict the same outcome obtained by Rosch.

Thus, as with perceptual categories, the assumption of the decomposability of category exemplars has not been a major issue. Most researchers assume the existence of semantic features. The major point of contention appears to be whether there are criterial features and/or logical rules for determining category membership.

Present Research

In the present research we used perceptual stimuli to test the claims of models of perceptual and semantic categorization, hence making the assumption that the same structural and processing principles apply to perceptual and conceptual categorization. We used schematic faces with well-defined features as stimuli corresponding to the assumption of the decomposability of words into semantic features.

Specifically, we were interested first in testing the Smith et al. model of categorization by using categories having both defining and characteristic features. We wanted to determine whether the typicality effects found in natural categories (effects on rate of learning, reaction times for verification, and subjective judgments of typicality) could be accounted for on the basis of the number of characteristic features of the members of artificial categories that have defining features.

Second, we were interested in determining which of several possible models best predicts reaction time for categorization, typicality judgments, and rate of learning for stimuli having a family resemblance structure. As noted previously, Reed (1972) has already compared several models. However, his work was completed before the results on typicality effects in natural language categories had been reported. Consequently, he did not obtain any judgments of typicality or record reaction times for verification of category membership. Although Hyman and Frost (1975) compared several models, they used stimuli that were not decomposable. Medin and Schaffer did compare their model to models similar to those presented by Hyman and Frost and did use stimuli with decomposable features; however, the categories they used were extremely small. Given that their model seems intuitively less plausible when categories are large, it would seem necessary to test it with larger categories. Therefore, in the present experiments fairly large categories were used for testing the various models.

Experiment 1

Because many models of categorization depend on distances among exemplars in a multidimensional space for their predictions, a multidimensional scaling study was done on the stimuli that would serve as exemplars in the categorization experiments. For both defining features and family resemblance categories, faces composed of

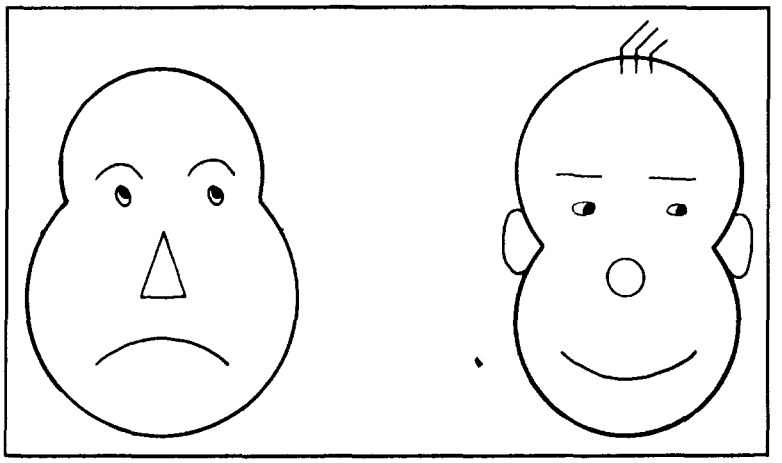

Figure 1. Faces showing examples of all the features used in Experiments 1–3.

seven features were used as exemplars. Examples of all the features are shown in Figure 1.

For the defining features condition, the categories were structured so that a conjunctive rule would define category membership: The members of Category 1, referred to as Harrys, had round noses and hair, whereas the members of Category 2, referred to as Charlies, had frowns and eyes looking to the left. The remaining features were more or less typical of one category or the other. For example, of the 12 Harrys, 9 had ears and 7 had thin faces, whereas of the 12 Charlies, 9 had no ears and 7 had fat faces. In order that a conjunctive rule would be necessary for determining category membership, some members of the contrasting category had one of the defining features of the other category (but never both). For example, 4 of the Charlies had round noses and 4 had hair, but none had both a round nose and hair. The exact structure of the categories is shown in Table 1.

For the family resemblance condition, the categories were structured so that a member of a category would have more feature overlap with the category it belonged to than with the other category. Feature overlap was computed by counting the number of times a feature of an exemplar was possessed by members of a category and summing these counts across all features of the exemplar. The structure of the categories and the family resemblance counts for the exemplars are shown in Table 2.

In this experiment, in each condition judgments of overall similarity of all pairs of the faces in that condition were obtained. A multidimensional scaling analysis was performed on the similarity data.

Method

Subjects

Twenty-four Johns Hopkins University students participated in this experiment; 12 in the defining features condition and 12 in the family resemblance condition. They were each paid $2 an hour for their participation.

Procedure

All 276 pairs of faces for each condition were presented in a random order through a Kodak Carousel slide projector. Each pair of faces was visible for 7 sec. Subjects were asked to judge the similarity of a pair on a scale from 1 (a very similar pair) to 9 (a very dissimilar pair). Subjects were asked to make their judgments quickly, based on the overall similarity of the pair. They were instructed to use the entire scale. Subjects wrote their responses on a sheet numbered from 1 to 276.

Results

Defining Features Condition

Before performing the multidimensional scaling analysis, the possibility of individual differences in similarity judgments was investigated by use of the inverse principal-

components analysis (similar to that employed by Tucker & Messick, 1963). This analysis treats subjects as variables in a principal-components factor analysis. Pairs of stimuli serve as the cases. For the uncentered matrix of stimulus pair by subject, the first principal component accounted for 58% of the variance, indicating a good degree of consistency across subjects. Centering this matrix by rows removes the degree to which subjects are behaving similarly as a total group and emphasizes differences across subjects. After centering, no distinct subgroups of subjects were evident. Thus, responses were averaged across all subjects. The average similarity data were analyzed by the KYST-2 multidimensional scaling program (Kruskal, Young, & Seery, Note 1). Solutions of varying dimensionality were tried, but a three-dimensional solution appeared adequate to account for the obtained similarity judgments. Stress (Kruskal, 1964) for the three-dimensional solution was .06, indicating a close monotonic fit of the three-dimensional configuration to the data.

The original configuration of the three-dimensional solution is shown in Figure 2. The first dimension appeared to be smile versus frown and the second, fat face versus thin face. These features were not, however, the only features affecting the position of the faces in the first two dimensions. If they had been, the faces would have clustered into four tight groups based on the four possible combinations of two values for

Table 1
Structure of Defining Features Categories

Category/ picture number	Hair[a]	Face shape[b]	Eyes[c]	Eye-brows[d]	Nose[e]	Mouth[f]	Ears[g]
Harry							
1	1	2	1	1	1	1	1
2	1	1	2	1	1	1	1
3	1	1	1	2	1	1	1
4	1	2	2	2	1	1	1
5	1	1	2	2	1	1	1
6	1	1	2	2	1	1	2
7	1	2	1	1	1	2	1
8	1	1	1	2	1	2	1
9	1	2	1	2	1	2	1
10	1	1	1	1	1	2	1
11	1	2	1	2	1	2	2
12	1	1	1	2	1	2	2
No. of 1s	12	7	8	4	12	6	9
Charlie							
1	2	2	2	2	1	2	2
2	2	1	2	2	1	2	2
3	2	2	2	1	1	2	1
4	2	1	2	1	1	2	2
5	2	2	2	1	2	2	2
6	1	2	2	2	2	2	2
7	2	1	2	1	2	2	2
8	1	2	2	1	2	2	2
9	2	2	2	2	2	2	2
10	1	1	2	1	2	2	2
11	2	2	2	1	2	2	1
12	1	1	2	1	2	2	1
No. of 1s	4	5	0	8	4	0	3

[a] 1 = hair, 2 = no hair. [b] 1 = thin face, 2 = fat face. [c] 1 = eyes left, 2 = eyes up. [d] 1 = curved eyebrows, 2 = straight eyebrows. [e] 1 = round nose, 2 = triangular nose. [f] 1 = smile, 2 = frown. [g] 1 = ears, 2 = no ears.

mouth shape and two values for face shape. Other features affected the relationship among faces in the plane defined by the first two dimensions but not in any systematic manner that would allow for dimensions corresponding to other features to be located. On the plane defined by the second and third dimensions, the faces clustered into groups that shared many features.

Family Resemblance Condition

The inverse principal-components method was again used to investigate the possibility of individual differences. For the uncentered matrix, the first principal component accounted for 46% of the variance, indicating that although subjects were not as similar as a total group for this condition as they were for the defining features condition, there was still a sizable degree of consistency across subjects. For the centered matrix, no distinct subgroups appeared; therefore the responses were averaged across subjects. Using the KYST-2 multidimensional scaling program, a three-dimensional configuration was found to be adequate, having a stress of .06. The three-dimensional configuration is shown in Figure 3. Again, the first two dimensions corresponded to smile versus frown and fat face versus thin face. A slight rotation of the third dimension would result in a comparison of triangular versus round nose. Again, these were not the only features affecting position on these dimensions.

Table 2
Original Family Resemblance Structure

Category/ picture number	Hair[a]	Face shape[b]	Eyes[c]	Eye-brows[d]	Nose[e]	Mouth[f]	Ears[g]	Family resemblance: To own category	Family resemblance: To contrasting category
Harry									
1	1	2	1	1	1	1	1	51	33
2	2	1	2	1	1	1	1	43	41
3	1	1	1	2	2	2	1	49	35
4	1	2	2	2	2	1	1	45	39
5	2	1	1	2	1	1	1	53	31
6	1	1	2	2	1	1	2	45	39
7	1	2	1	1	1	2	1	47	37
8	1	1	1	2	2	1	1	53	31
9	2	2	1	2	2	1	1	47	37
10	2	1	1	1	1	1	1	49	35
11	1	2	1	2	1	2	2	45	39
12	1	1	1	2	1	2	2	47	37
No. of 1s	8	7	9	4	8	8	9		
Charlie									
1	2	2	2	2	1	2	2	49	35
2	2	1	2	2	1	2	2	47	37
3	2	2	2	1	1	2	1	47	37
4	2	1	1	1	1	2	2	45	39
5	2	2	2	1	2	1	2	53	31
6	1	2	2	2	2	1	2	45	39
7	2	1	2	1	2	1	2	51	33
8	1	2	2	1	2	2	2	53	31
9	2	2	1	2	2	1	2	43	41
10	1	1	1	1	2	2	2	45	39
11	2	2	2	1	2	2	1	51	33
12	1	1	2	1	2	2	1	45	39
No. of 1s	4	5	3	8	4	4	3		

[a] 1 = hair, 2 = no hair. [b] 1 = thin face, 2 = fat face. [c] 1 = eyes left, 2 = eyes up. [d] 1 = curved eyebrows, 2 = straight eyebrows. [e] 1 = round nose, 2 = triangular nose. [f] 1 = smile, 2 = frown. [g] 1 = ears, 2 = no ears.

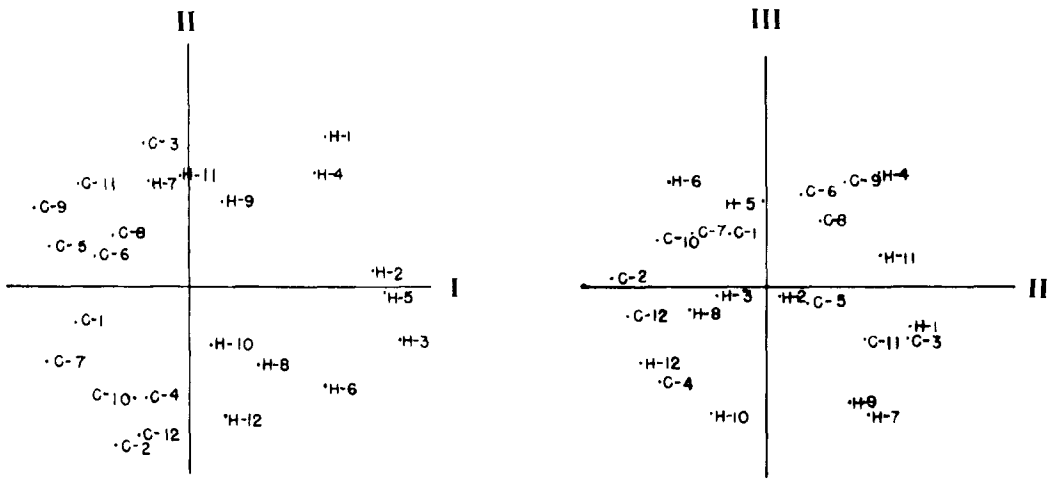

Figure 2. Three-dimensional multidimensional scaling solution for the defining features condition. (C = Charlie; H = Harry.)

Faces sharing the same values of face shape, mouth, and nose could be quite distant from each other, but again the dispersions could not be traced in any systematic manner to other features.

Experiment 2

In this experiment, subjects were required to learn the category memberships of the exemplars of the defining features categories (Table 1). We were interested in recording the number of trials necessary to learn the correct classification of each face and reaction times for classification once category membership had been learned and in obtaining typicality judgments. The model to be tested was that of Smith et al., which assumes that the number of characteristic features of an exemplar affects categorization time and the judged typicality of the exemplar.

In Table 3, column 1 is the number of characteristic features for each face. Be-

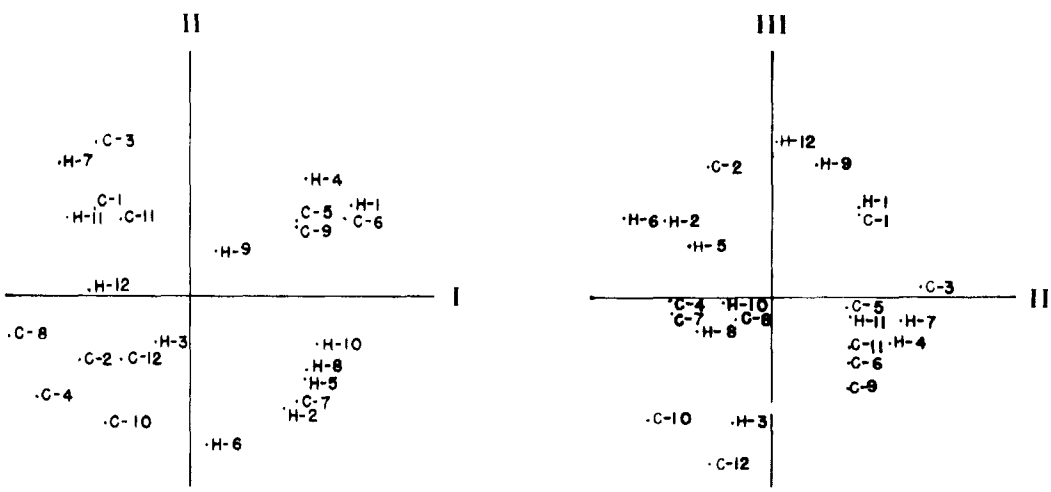

Figure 3. Three-dimensional multidimensional scaling solution for the family resemblance condition. (C = Charlie; H = Harry.)

cause all of the features were binary, if more than 6 of the faces in a category had a particular feature, that feature was considered to be a characteristic feature. (Because 6 Harrys had smiles and 6 had frowns, neither feature value for mouth was considered to be characteristic of the Harrys.) The second column shows a weighted value for number of characteristic features; each feature has been weighted by the degree to which it is characteristic of the category. For example, since 9 of the 12 Harrys have ears and 7 of the 12 have thin faces, having ears was given a weight of 9 and having a thin face was given a weight of 7. Besides the degree to which it is characteristic of a category, the perceptual salience of a feature may also play a role in determining the importance of a feature in relation to typicality. Thus, another weighted score was computed in which the features were weighted by their relative salience in the scaling solution. Since smile–frown accounted for 46% of the variance in the scaling solution, and face shape for 34%, these features were assigned weights of 4 and 3, respectively. All other features were given weights of 1.

Table 3
Defining Features

Category/ picture number	No. of characteristic features	Characteristic features: Weighted 1	Characteristic features: Weighted 2	Average exemplar	Nearest exemplar	Prototype	Medin & Schaffer (1978)
Harry							
1	2	56	60	.479	.173	.591	.984
2	2	54	62	.604	.818	.783	.975
3	4	62	70	.612	.840	.711	.993
4	2	56	60	.376	.341	.443	.976
5	3	58	66	.562	.746	.637	.987
6	2	52	60	.330	.381	.377	.966
7	2	56	60	−.024	.088	−.018	.886
8	4	62	70	.206	.246	.351	.930
9	3	60	64	.215	.185	.296	.917
10	3	58	66	.104	.020	.016	.878
11	2	51	58	−.053	−.259	.010	.797
12	3	56	64	.068	−.006	.128	.849
Correlation with							
TLE	−.02	−.13	−.19	−.93	−.88	−.86	−.91
RT	.23	.16	.09	−.85	−.74	−.82	−.68
Typicality judgments	−.13	−.05	.05	.92	.84	.89	.85
Charlie							
1	3	60	50	.601	.572	.729	.918
2	2	58	44	.459	.397	.404	.906
3	3	58	48	.139	.041	.110	.891
4	3	62	48	.235	.084	.201	.941
5	5	60	50	.561	.331	.644	.987
6	3	52	42	.507	.208	.514	.886
7	4	58	44	.670	.761	.726	.981
8	4	56	46	.425	.235	.433	.924
9	4	56	46	.610	.364	.584	.979
10	3	54	40	.425	.344	.403	.904
11	4	54	44	.373	.097	.367	.969
12	2	48	34	.350	.121	.293	.815
Correlation with							
TLE	−.80	.10	−.16	−.59	−.18	−.58	−.63
RT	−.66	−.22	−.40	−.47	−.21	−.48	−.69
Typicality judgments	.73	.17	.50	.58	.26	.60	.53

Note. TLE = trial of last error; RT = reaction time.

Method

Subjects

Twelve Johns Hopkins University students were each paid $2.50 per hour to participate in this experiment.

Procedure

Categorization task. Subjects were told that their task was to learn the category membership of 24 faces. They were given no indication as to the type of category structure used. They were told that they would be given feedback on their classifications that would help them learn category membership. Slides of the faces were projected on a rear-projection screen. Subjects initiated each trial by pressing a button. After a 1-sec interval a face appeared on the screen and remained there for 3 sec. Subjects were required to indicate which category the face belonged to by pressing one of two buttons labeled either "Harry" or "Charlie." After the subject responded the experimenter would tell him or her whether the response was correct or incorrect. All 24 faces were presented in a random order. Nine different random orders were used. Trials continued until a subject completed two sets of the 24 faces without making any errors.

After reaching this criterion, five more sets of the 24 faces were presented, during which reaction times for categorization were recorded. Subjects were instructed to respond as quickly as possible without making any errors.

Typicality judgments. Following the reaction time trials, subjects were asked to judge how typical each face was of the category it belonged to. Subjects were presented with all of the faces in one category, one at a time, and then with all of the faces in the other category. The order in which the categories were presented was counterbalanced across subjects. Subjects were instructed to use a scale from 1 (a very typical exemplar) to 9 (a very atypical exemplar).

Similarity judgments. Subjects were presented with all pairs of the 24 faces, which had been reproduced on 8½-inch × 11-inch (21.6 cm × 27.9 cm) paper. They were asked to judge the similarity of each pair on a scale of 1 (very similar faces) to 9 (very dissimilar faces). They were told to make their judgments quickly and to use the entire scale.

Subjective reports. At the end of the experimental sessions subjects were asked to describe the basis of their typicality judgments.

Results

Learning

The trial on which the last error occurred (TLE) was determined for each face for each subject, and averages were computed. The correlations of these averages with the various characteristic feature scores are shown in Table 3. For the Harry category, very low correlations were obtained for all three scores. In the Charlie category, the unweighted characteristic feature score had a high correlation with trial of last error, but the weighted scores had very low correlations.

An analysis of variance was performed on the learning data. The individual faces were treated as a nested factor within categories. There was no significant difference between categories, but there was a significant difference among faces within a category, $F(22, 242) = 2.04$, $p < .01$. A post hoc comparison of the Harrys having a smile to those having a frown was significant, $F(1, 242) = 39.9$, $p < .01$, accounting for 65% of the variance among the means. Within the Charlie category, the comparison between those having a round nose and those having a triangular nose was also significant, $F(1, 242) = 8.8$, $p < .01$, accounting for 13.6% of the variance. Within the Charlies having a triangular nose, the comparison of those having hair to those without hair was significant, $F(1, 242) = 5.7$, $p < .05$, accounting for 8.7% of the variance. Another feature that seemed to have a marginal effect on learning was face shape. Across both categories, the comparison of the faces having a typical face shape to those having an atypical face shape accounted for 10.4% of the variance. However, this comparison was not orthogonal to those discussed previously.

Reaction Times and Typicality Judgments

Mean reaction times for the 24 faces were computed. Table 3 shows the correlation between these means and the various characteristic feature scores. Very low correlations were obtained on all three measures for the Harry category, and a moderate correlation was obtained for the Charlie category only for the number of characteristic features (unweighted).

Since it is possible that subjects did not learn the categories in terms of defining features but rather treated them as examples of family resemblance categories, and since the nondefining features were not irrelevant to categorization but partially predictive of category membership, it was possible to work out the predictions of the various family resemblance models for this stimulus

set. The Rosch model predicts that it is degree of feature overlap that determines reaction time for categorization. The predictions of this model would correlate perfectly with the first weighted characteristic feature score, which was found to have very low correlations with reaction time for both the Harry and Charlie categories.

Of the models proposed by Hyman and Frost (1975), the exemplar models and the prototype model can be tested in the present experiment. The rule model would predict equal reaction times for all stimuli, since all category members are equally good examples of the rule defining category membership. Hyman and Frost's average exemplar model, however, predicts reaction time to be a function of the average distance of the exemplar to the members of its category minus the average distance of the exemplar to the exemplars of the contrasting category. Column 4 in Table 3 shows these values computed from distance in the multidimensional space. The nearest exemplar model predicts reaction time to be a function of the distance of an instance to the nearest exemplar in the instance's category minus the distance of the instance to the nearest exemplar in the contrasting category. These values are shown in Table 3, column 5. The prototype model predicts that distance to the category prototype minus distance to the prototype of the contrasting category determines reaction time. Prototypes for each category could not be determined on the basis of average feature values, since discrete features were used. Instead, the centroid of each category in the multidimensional space was computed, and distance from this centroid was regarded as distance to the prototype. These distances are shown in column 6 of Table 3.

In order to compute the predictions of the Medin and Schaffer model, it was necessary to assume similarity parameters for the seven features. Their model assumes that the similarity of two values of a feature can be represented by a parameter ranging in value from 0 to 1 (where 1 is identity). The value of this parameter is related to the discriminability and salience of the feature values. Overall similarity of two exemplars is determined by multiplying together the similarity parameters for each feature of the exemplars. The function determining overall evidence favoring classifying stimulus i in Category A is assumed to be the summed similarity of the instance to all stored exemplars of Category A divided by the summed similarity of the instance to all exemplars in the present context (all members of Category A plus all members of Category B when there are two possible categories). Reaction time for categorization is assumed to be a decreasing function of the amount of evidence favoring classification.

Medin and Schaffer determined the similarity parameters for their stimuli after the fact by finding those weights that maximized the correlation between predicted and observed results for both their model and the competing models they were testing. Because the weights for the characteristic feature score and the distances for the Hyman and Frost models were determined by the results of the multidimensional scaling analysis of these stimuli, the similarity parameters for the Medin and Schaffer model were also assigned on the basis of the scaling solution. Because the dimensions of smile–frown and face shape accounted for the most differentiation between stimuli in the scaling solution, these features were assigned similarity parameters of .1. The remaining features were assigned parameters of .5. Evidence favoring classification in its own category was computed for each face in terms of the context model, and these values are shown in column 7, Table 3.

For the Harry category, all of the Hyman and Frost models and the Medin and Schaffer model had moderate correlations with reaction times (as shown in Table 3), with the average exemplar model having the highest correlation ($r = .85$, $p < .01$). For the Charlie category, the average exemplar and prototype models of Hyman and Frost had moderate correlations with reaction times but just missed significance at the .05 level. The Medin and Schaffer model was most highly correlated with reaction times for the Charlie category.

Mean typicality ratings were also computed and correlated with the same models. Except for the fact that the correlations

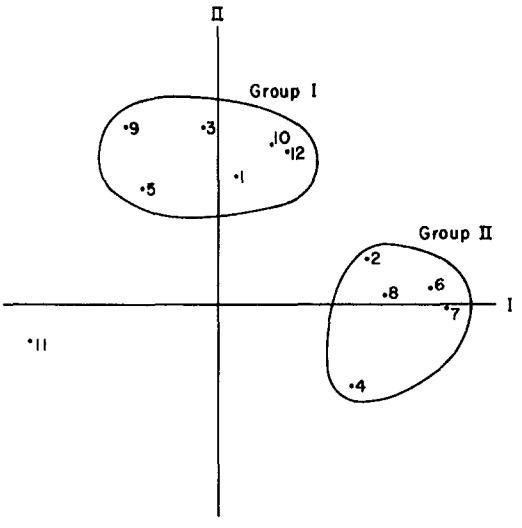

Figure 4. Inverse principal-components analysis of reaction times for subjects in Experiment 2.

have opposite signs, the correlations for the typicality judgments followed the pattern obtained for the reaction times. Average trial of last error was also correlated with these models. For the Harry category, the Medin and Schaffer model and all of the Hyman and Frost models had high correlations with trial of last error. For the Charlie category, the pattern of correlations for trial of last error was like that obtained for the reaction times and typicalities, with moderate correlations being obtained for the Medin and Schaffer models and the average exemplar and prototype models of Hyman and Frost.

Although some of the models had high correlations with reaction times for one category, none did a good job of predicting reaction times for both categories. In order to investigate further the pattern of reaction times, an analysis of variance was performed on the reaction times. The between-categories difference and the within-category differences failed to reach significance: between categories, $F(1, 11) = 4.4, p < .10$; within categories, $F(22, 242) = 1.6, p < .10$. However, subjects' informal remarks about how they performed the categorization task indicated that subjects had been using very different strategies. Analyzing together groups of subjects who performed the task differently may have resulted in the lack of significant differences.

In order to evaluate individual differences quantitatively, the inverse principal-components analysis, usually reserved for similarity data, was applied to the reaction time data. The first two principal components of this analysis are shown in Figure 4. Two groups of subjects appeared in this analysis, one containing six subjects and the other containing five. The remaining subject was well separated from the rest of the subjects and was eliminated from further analysis. The group data were analyzed by analysis of variance. For Group 1 there was no significant difference between categories ($F \simeq 1$) and a significant difference within categories, $F(22, 110) = 2.61, p < .01$. Within the Harry category, there was a significant difference between subject reaction times for faces with a smile and for those with a frown, $F(1, 110) = 4.9, p < .05$, accounting for 8.5% of the variance between the means for the faces. Within the Charlie category, there was a significant difference between the faces having a round nose and those having a triangular nose, $F(1, 110) = 13.2, p < .01$, and, among those having a triangular nose, a significant difference between those having hair and those without hair, $F(1, 110) = 24.3, p < .01$. These two comparisons within the Charlie category accounted for 65% of the overall variance between the means. The remaining orthogonal comparisons within each category failed to reach significance.

For Group 2 there was a significant difference between categories, $F(1, 4) = 23.1, p < .01$, and no significant difference within categories ($F \simeq 1$). The average reaction times were 669.5 msec for the Harry category and 869.8 msec for the Charlie category.

From subjects' comments it appeared that the subjects in Group 2 learned the defining features for the Harry category, the round nose and hair, and categorized the Charlies by default. They used a combination of features to determine membership in the Harry category (Hayes-Roth & Hayes-Roth, 1977) and assigned an exemplar to the Charlie category after deciding that it was not a Harry. Given that there

were no significant differences within categories, the results of the analysis of variance support this explanation. The subjects in Group 1 took a different approach. They noticed that if a face had a triangular nose or the head were bald, then it must be in the Charlie category. Thus, one would expect longer reaction times for the Charlies with round noses or with hair, and the results of the post hoc comparisons indicated significant effects for the presence of these features on reaction times. The significant difference between smiling faces and frowning faces in the Harry category is unexpected, since subjects using the Group 1 strategy would not have had to use information about the mouth in order to categorize the Harrys. This small but significant difference could have resulted because some subjects used the mouth as part of a strategy. For instance, subjects could have noticed that the triangular nose indicated Charlie but used a feature other than hair (such as the mouth) for discriminating the round-nosed Charlies from the round-nosed Harrys. On the other hand, this difference between smiling and frowning Harrys could indicate the effect of a salient perceptual feature on reaction time even when this feature is not necessary for categorization. This interpretation of the difference between smiling and frowning Harrys would indicate support for the Smith et al. proposal, but only weak support, since only one of the characteristic features had such an effect. Moreover, the feature was not truly characteristic of the Harry category, since half had smiles and half had frowns. Smile was characteristic of the Harry category only in that it was totally absent from the Charlie category. It shold be noted that for Group 2, which used the defining features of the Harry category, there was also a trend for the smiling Harrys to be classified faster than the frowning Harrys (mean reaction time was 630.2 msec for the smiling Harrys and 708.8 msec for the frowning Harrys), but as noted previously, differences between means for the faces within categories were far from being significant for Group 2.

An analysis of variance was also performed on the typicality judgments. For all subjects combined there was a significant difference between typicalities within each category. Within the Harry category, 93% of the difference between the mean typicality ratings could be accounted for by smile versus frown, with the smiling faces judged to be more typical. Within the Charlie category, 67% of the variance was accounted for by face shape. Obtained typicalities correlated −.91 with reaction times within the Harry category, whereas within the Charlie category the correlation was −.60.

The high correlation between the obtained typicality values and the reaction times for the Harrys can be accounted for almost entirely by the smile–frown difference in the typicality values. If the obtained typicalities for the smiling Harrys were replaced with 1s and the typicalities for the frowning Harrys replaced with 2s, the correlation with reaction times would still be −.89.

The correlation for the Charlie category between typicality judgments and reaction times can best be understood by looking at the typicalities by groups of subjects found in the analysis of the reaction times. For Group 1—the group that used baldness and triangular nose for defining the Charlies—there were significant differences between typicality judgments for both the Harry and Charlie categories. Within the Harry category, 91% of the variance could be accounted for by the distinction between smiling and frowning faces. Within the Charlie category, 46% of the variance could be accounted for by the comparison of those with triangular noses to those with round noses, with the triangular-nosed Charlies being judged more typical, and 37% of the variance could be accounted for by the comparison of those with hair to those without hair. For Group 2—the group that learned the defining features of the Harry category—significant differences in typicality judgments were also found among the faces within each category. For the Harry category, 91% of the variance could be accounted for by the comparison of those with smiles to those with frowns. Within the Charlie category, the only significant comparison based on two values of a feature was

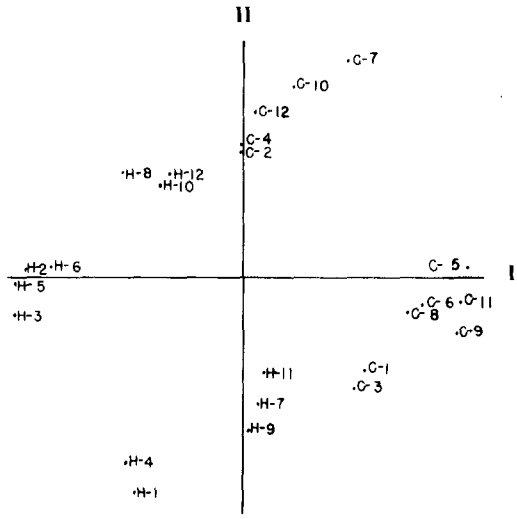

Figure 5. First two dimensions of multidimensional scaling solution for defining features condition after categorization. (C = Charlie; H = Harry.)

that between those having thin faces and those having fat faces, accounting for 52.4% of the variance. When combining the groups, the reaction times for the Charlies would, at most, reflect the differences found for Group 1, since there were no significant differences between the mean reaction times for Group 2. However, when combining the two groups' typicality judgments, both the effects found for Group 1 and those found for Group 2 would be reflected. Hence, there was some correlation (−.60) between combined reaction times and typicalities, but not as great a correlation as that found for the Harry category.

Similarities

The similarity judgments of the subjects in the categorization task were analyzed in the same manner as those of the subjects in Experiment 1. An inverse principal-components analysis was first performed. For the uncentered matrix of stimulus pairs by subjects, the first principal component accounted for 53% of the variance, indicating a good deal of consistency across subjects. In the analysis of the centered matrix, no obvious groupings of subjects were found. Thus, although groups of subjects could be identified in the analysis of reaction times obtained in the categorization task, groups of subjects were not evident in the analysis of similarity judgments. One might have expected that the features used in a group's categorization strategy would influence the weight given a feature in their similarity judgments. However, since both groups of subjects used the same features in their tests (hair and nose), the similarity judgments of the two groups would not be expected to differ.

The similarity data were averaged across all subjects and analyzed by the KYST-2 multidimensional scaling program. The four-dimensional solution had a stress of .07, and the five-dimensional solution, a stress of .04. Since all five dimensions were interpretable, the five-dimensional solution was chosen. In the plot of the first two dimensions shown in Figure 5, the faces segregated into two groups on the basis of face shape. The other dimension in this plane perpendicular to that for face shape corresponded both to smile versus frown and to triangular versus round nose, that is, the faces were ordered from those having smiles (on the left) to those having frowns (on the right). However, within those having frowns, those having triangular noses were farthest to the right, and those having round noses were closer to the members of the Harry category (farthest to the left). The third dimension corresponded to eyes looking left versus eyes looking up, the fourth dimension to those with hair versus those without hair, and the fifth to those with straight eyebrows versus those with curved eyebrows. Thus, the only feature not represented was that of ears versus no ears. It appeared that subjects who had participated in the categorization task were more analytical in their judgments; that is, they based their similarity judgments on the individual features of each face rather than making more holistic judgments. Also, a feature that had been important in the categorization task, shape of nose, was found to play an important role in the similarity judgments, being correlated with the dimension of smile–frown. This feature had not appeared in the plane of the first two dimensions for the scaling solution obtained in Experiment 1 for the defining features

case. There also appeared to be some effect for category membership per se in the plane of the first two dimensions. Unlike Experiment 1, a line could be drawn in the first two dimensions. Unlike Experiment 1, a line could be drawn in this plane separating the Harrys and Charlies. However, an even greater separation exists in this graph between those faces having a smile and those having a frown than exists between the two categories.

Discussion

The purpose of Experiment 2 was to test whether nondefining features could have an effect on reaction times and on typicality judgments. For the subjects who did learn the defining features of one of the categories (Group 2), there was no significant difference between the means for the different faces within a category. Typicality judgments, however, were related to characteristic features of the faces. For the Harry category, the smiling Harrys were judged to be more typical of the category than the frowning Harrys. Although neither type of mouth was actually typical of the Harry category, the smile made the face more distinct from the Charlie category. Rosch and Mervis (1975) obtained a similar effect, finding that when controlling for degree of overlap within a category, the stimuli having less overlap with the contrasting category were judged to be more typical of their own category. Within the Charlie category, the salient feature of face shape was found to influence typicality judgments. These results for typicality judgments may in fact indicate nothing about how subjects organize categories; they may solely represent subjects' attempts to make reasonable responses in the task assigned to them. Having noticed during the learning phase that certain features were more representative of a category, subjects could have used this knowledge in making their typicality judgments. Salient features had an effect because these were the features subjects would have been most likely to notice.

For the group that apparently did not learn the conjunctive rule for determining category membership (Group 1), differences between mean reaction times for the different faces within a category were found to be significant. However, a large percentage of the variance between the means for the faces can be accounted for on the basis of the comparisons for the two features used to classify the faces (triangular nose and baldness) within the Charlie category. These results indicate an effect for features that, for these subjects, were defining rather than characteristic. The significant difference between the smiling and frowning Harrys is the only evidence from this group that may indicate the effect of a nondefining feature on reaction times. The results for the typicality judgments for this group corresponded to those obtained for the reaction times: Significant differences among the means for the Charlies resulted from comparisons between nose shapes and from the presence or absence of hair. Also, a significant effect was found for smile–frown within the Harry category for the typicality judgments even though subjects appeared not to have used this feature in making their decisions about category membership.

An overall consideration of the data we have reported suggests that the Smith et al. model received, at best, little support from these findings. However, it could be argued that the Smith et al. model predicts an effect of characteristic features on reaction time only when some categorization decisions can be made during the first stage of the model. In a situation in which the contrasting categories are very similar, subjects would have to set their criterion for first stage decisions very high in order to avoid making errors. If this criterion were so high that essentially all decisions were made at the second stage (where defining features are checked), then one would expect no effect of characteristic features on reaction time.

However, the overlap between the categories used in this experiment was not so great that subjects should have been induced to use second stage processing for all decisions. If one interpreted the Smith et al. model to imply that in the first stage of processing an exemplar is compared to a category representation consisting of the defining features plus all the characteristic

features, then even the most atypical member of the categories used here would be more similar to its own category representation than to that of the contrasting category. The most atypical members had four features (two defining and two characteristic) in common with their own category representation and three in common with the contrasting category.

If instead one were to assume that the first stage is used to determine global similarity rather than number of feature matches, there is only one face that was more similar to the prototype of the contrasting category than it was to the prototype of its own category, as is shown in column 4 of Table 5. Looking at the values in this column, one can see that subjects could have made many decisions at the first stage. For example, if the criterion for a Stage 1 decision were .250, then 8 of the 12 Harrys and 10 of the 12 Charlies could have been classified at Stage 1.

In sum, whether one assumes the first stage to consist of a count of feature matches or of a determination of similarity to a prototype, this stimulus set allowed for many Stage 1 decisions. The lack of correlation between reaction times and the predictions of the Smith et al. model indicates that subjects did not use a process like the first stage of this model.

The relative success of the other models, in terms of correlations between the models' predictions and the reaction times, typicality judgments, and learning data can be understood on the basis of the importance given to the similarity of exemplars of one category to those of the contrasting category and the weights given to the features by a particular model. For the Harry category, the only significant difference in reaction times between stimulus means was that between the faces having a smile and those having a frown found for Group 1. Because each of the two mouths was equally represented in this category, this feature had no effect on the characteristic feature score or on any of the weighted characteristic feature values. Hence, low correlations were obtained with predictions based on characteristic feature models (Rosch & Mervis, 1975; Smith et al., 1974). The models of Hyman and Frost and that of Medin and Schaffer take into account not only the similarity of an exemplar to its own category but also its similarity to the contrasting category. Thus, the smiling Harrys were predicted to be classified faster than the frowning Harrys for these models, since the smiling Harrys were more dissimilar to the Charlie category. Hence, high correlations were obtained for these models. Similarly, for the typicality judgments and trial of last error data, the comparison accounting for most of the variance between means was that for the faces having a smile versus those having a frown. Consequently, the same models had high correlations with these measures.

For the Charlie category, the characteristic feature models fared somewhat better. Significant differences in reaction times were obtained between the faces with a triangular nose and those having a round nose and between those having hair and those that were bald. Because triangular nose and baldness were characteristic features for this category, these features influenced the value of the characteristic feature score; therefore, the moderate correlations between these values and reaction times would be expected. The lower correlations for the weighted characteristic feature scores resulted from the low weights given to nose and hair in computing these values. Similarly, the predictions of the Hyman and Frost models had low correlations with reaction by face shape and smile–frown and not by the nose and hair features. The Medin and Schaffer model was somewhat more successful in predicting reaction times, because although the similarity parameters were biased toward lending more influence to face shape and smile–frown, the fact that baldness and triangular nose were totally uncharacteristic of the Harry category resulted in the Medin and Schaffer model's predicting faster reaction times for faces having these features. Differences among the mean typicality judgments and average trial of last error for the Charlie category were also found to depend mainly on the contrasts between round nose–triangular nose and hair–no hair. The correlations between these measures and the predictions of the various models followed

the same pattern as that obtained for the reaction times.

However, the major conclusion to be drawn from this experiment is that for these categories with defining features, subjects' behavior was like that found in the traditional concept formation studies: Subjects discovered logical rules for classifying the exemplars. For one group, the rule used corresponded to a conjunctive rule, that is, the presence of the two criterial features for the Harry category (hair and round nose). This group apparently did not discover the conjunctive rule used to define the Charlie category but instead classified as Charlies those faces not having the combination of features necessary to be Harrys. The second group discovered a more complex rule for the classification of the faces. Charlies, for this group, were those either having a triangular nose or having a round nose and being bald. Harrys were those faces that were not Charlies.

Thus, subjects in this experiment did not appear to use overall family resemblance or overall similarity to category members or category prototypes in deciding category membership. Even those subjects who did not discover the conjunctive rule for categorization intended by the authors developed another type of logical rule that could be used successfully. It should be noted that neither group discovered the defining features for both categories. With a finite set of categories, it is always possible that one category can be defined as that category which does not have the features of the other categories. For only two categories, the strategy of defining one category by default is an efficient one, since it requires that subjects learn only the characteristics of one category. As the number of categories becomes larger, this strategy would become less attractive. Given that there are n categories, in order to classify the members of the default category, features for the $n-1$ categories would have to be checked before coming to the correct classification decision. Since in natural settings there are a very large number of categories to choose from, experiments using a large number of categories would more closely approximate a real-life situation. However, there is no obvious reason to assume that subjects' behavior would cease to be rule-based in those cases where they were required to deal with more than two categories with defining features except that the default strategy ceases to be an optimal one.

Experiment 3

The third experiment was designed to test the various family resemblance models by using categories that had no defining features. The stimulus structure used is shown in Table 2. The two categories were symmetric with respect to each feature, that is, if one category had n faces with a certain feature, the other category had $12 - n$ faces with this feature. For example, 9 of the 12 Harrys had ears, so 3 of the Charlies had ears. The categories were also structured so that each face would have a greater family resemblance to its own category than to the other category, with family resemblance being determined in terms of feature overlap. Family resemblance scores for each face in its own category and in the contrasting category are also shown in Table 2.

Three subjects were tested with this stimulus set in the categorization task, using procedure and instructions identical to those in Experiment 2. However, none of the subjects learned the categories and were, after 1 hour of testing, still making six to eight errors on the 24 faces.

Two factors appeared to contribute to the difficulty of learning this category structure. First, the features most predictive of category membership (eyes and ears) were not the most salient features in the scaling solution; consequently, category members may have appeared too dissimilar in terms of salient perceptual features. Second, there was evidence in the pattern of errors that subjects focused on one feature at a time or a combination of two features and never gave up on this strategy even though it was not successful.

In the revised category structure (shown in Table 4), the most salient features (mouth and face shape) were made the most characteristic features. Also, the instructions were

revised so as to stress that no feature was possessed by all the members of a category.

Family resemblance scores for the revised structure are shown in Table 5. Predictions of the other models, those of Hyman and Frost and Medin and Schaffer, were worked out for this category structure in the same manner as they had been for the defining features categories. Distances in the multidimensional space were used to make predictions for the Hyman and Frost models. For the Medin and Schaffer model, similarity parameters were estimated on the basis of the importance of the features in the multidimensional scaling solution for these faces. Smile–frown and face shape were assigned parameters of .1; nose, a parameter of .3; and the remaining features, .5. The predictions of these models are also shown in Table 5.

Method

Subjects

Seventeen Johns Hopkins University students were each paid $2 per hour to participate in the experiment. None of these subjects had participated in any of the previous experiments.

Procedure

The procedure was like that used in Experiment 2 but with two exceptions. The instructions were revised as noted above, and following the typicality judgments, subjects were asked to describe in detail the method they had used to learn the categories and the basis of their typicality judgments.

Results

Learning Data

Of the 17 subjects, 8 failed to meet the learning criterion in 1 hour of testing and

Table 4
Structure of Revised Family Resemblance Categories

Category/ picture number	Hair[a]	Face shape[b]	Eyes[c]	Eye-brows[d]	Nose[e]	Mouth[f]	Ears[g]
Harry							
1	1	2	1	1	1	1	1
2	1	2	2	2	2	1	1
3	1	2	1	1	1	2	1
4	1	1	1	2	2	1	1
5	2	2	1	2	2	1	1
6	2	1	1	1	1	1	1
7	2	2	2	2	1	2	2
8	2	2	2	1	1	2	1
9	2	2	2	1	2	1	2
10	1	2	2	2	2	1	2
11	2	2	1	2	2	1	2
12	2	2	2	1	2	2	1
No. of 1s	5	2	6	6	5	8	8
Charlie							
1	2	1	2	1	1	1	1
2	1	1	1	2	2	2	1
3	2	1	2	2	1	1	1
4	1	1	2	2	1	1	2
5	1	2	1	2	1	2	2
6	1	1	1	2	1	2	2
7	2	1	2	2	1	2	2
8	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2
10	1	2	1	1	2	2	2
11	1	1	1	1	2	2	2
12	1	1	2	1	2	2	1
No. of 1s	7	10	6	6	7	4	4

[a] 1 = hair, 2 = no hair. [b] 1 = thin face, 2 = fat face. [c] 1 = eyes left, 2 = eyes up. [d] 1 = curved eyebrows, 2 = straight eyebrows. [e] 1 = round nose, 2 = triangular nose. [f] 1 = smile, 2 = frown. [g] 1 = ears, 2 = no ears.

consequently did not participate in the remaining tasks. Average trial of last error was computed for each face for the 9 subjects who met the learning criterion. For the Harry category, high correlations were obtained between these averages and the predictions of the Rosch and Medin and Schaffer models. For the Charlie category, the highest correlation obtained was for the Rosch model. The rest of the correlations were quite low for the Charlie category. These correlations are shown in Table 5.

An analysis of variance was performed on the trial of last error data. A significant difference was obtained both between categories, $F(1, 8) = 7.3$, $p < .05$, and among the faces within a category, $F(22, 176) = 3.98$, $p < .01$. The Harrys were learned sooner than the Charlies, with average TLE equal to 9.28 for the Harrys and 11.49 for the Charlies. A post hoc comparison of those faces having a characteristic face shape to those having an uncharacteristic face shape within the Harry category was significant, $F(1, 176) = 15.0$, $p < .01$, and accounted for 16.7% of the variance among the means for all the faces. Within the Harry category, the comparison of those having a smile to those having a frown (among those having the characteristic face shape) ac-

Table 5
Family Resemblance

Category/ picture number	Family resemblance	Average exemplar	Nearest exemplar	Prototype	Medin & Schaffer (1978)
Harry					
1	41	.483	.297	.510	.960
2	43	.569	.695	.622	.973
3	37	.181	−.015	.120	.806
4	35	.054	.183	.004	.815
5	45	.203	.012	.214	.976
6	35	.143	.060	.123	.558
7	35	.125	.050	.070	.828
8	39	.303	.110	.261	.952
9	41	.480	.479	.590	.920
10	39	.526	.582	.527	.946
11	41	.460	.568	.481	.969
12	41	.150	.105	.129	.948
Correlation with					
TLE	−.81	−.56	−.39	−.54	−.81
RT	−.75	−.59	−.46	−.61	−.86
Typicality judgments	.84	.65	.47	.64	.89
Charlie					
1	35	.155	.329	.152	.747
2	39	.193	.364	.226	.919
3	35	.047	.022	.034	.835
4	41	.352	.347	.428	.913
5	37	−.041	.325	.007	.805
6	45	.163	.153	.189	.974
7	43	.332	.348	.402	.917
8	43	.481	.550	.600	.946
9	37	.069	−.050	.109	.852
10	35	.375	.383	.447	.861
11	43	.433	.573	.489	.978
12	39	.288	.513	.286	.949
Correlation with					
TLE	−.59	−.08	.07	−.07	−.29
RT	−.63	−.43	−.05	−.41	−.53
Typicality judgments	.60	.36	.15	.29	.63

Note. TLE = trial of last error; RT = reaction time.

counted for 12.0% of the variance, $F(1, 176) = 10.5, p < .01$. In the Charlie category, the comparison for face shape accounted for 8.7% of the variance, $F(1, 176) = 7.63$, $p < .01$. Within the Charlie category (among those having a characteristic face shape), a significant difference was found between those having ears and those without ears, $F(1, 176) = 15.5$, $p < .01$, accounting for 18.8% of the variance. Among those Charlies that had no ears (Charlies 1, 2, 3, and 12), there was a significant difference between Charlie 2 and the rest, $F(1, 176) = 7.0$, $p < .01$, accounting for 8.0% of the variance. The average trial of last error for Charlie 2 was 17.11, whereas the average for Charlies 1, 3, and 12 was 11.11.

A brief analysis was made of the learning data for the eight subjects who failed to reach criterion. Since there were no last errors for several faces for each of these subjects, number of errors for each face was computed. This data revealed that the four faces having the highest number of errors were those having uncharacteristic face shapes (9.9 mean errors for the 4 faces with uncharacteristic face shapes vs. 4.4 mean errors for the remaining 20 faces). Inspection of the pattern of errors for individual subjects showed that six subjects on one or more trials missed only faces having an uncharacteristic face shape, indicating that face shape alone may have been used at one point for classifying the faces. The other two subjects never showed this pattern of errors. One of these apparently focused on a combination of nose and mouth and then switched to ears and mouth. The pattern of errors for the remaining subject resisted interpretation.

Reaction Times and Typicality Judgments

Correlations of the mean reaction times and the various models are shown in Table 5. For the Harry category, high correlations were obtained between the reaction times and the family resemblance scores computed according to Rosch's model and with the predictions of Medin and Schaffer's context model. Only moderate correlations were obtained between the reaction times and any of the Hyman and Frost models. For the Charlie category, moderate correlations were obtained for the family resemblance scores and the context model, but again, low correlations were obtained for the Hyman and Frost models. The same pattern of results was obtained when correlating mean typicalities with the predictions of the various models.

The reaction time data were analyzed by analysis of variance. There was no significant difference between the reaction times for the two categories, $F(1, 8) = 1.89, p > .10$, but there was a significant difference within categories, $F(22, 176) = 5.02$, $p < .01$. The comparison between the faces with the typical face shape for their category and those with the odd face shape for their category was highly significant, $F(1, 176) = 76.2$, $p < .01$, and accounted for 70% of the variance of the means. Within the Harry category, no other comparisons were significant. Within the Charlie category, among those with the characteristic face shape for the category, there was a significant difference between the faces having ears and those without ears, $F(1, 176) = 16.6$, $p < .01$, accounting for 15% of the total variance among the means for the faces.

Because subjects were asked to describe the manner in which they had learned to which category the faces belonged, we were fairly confident that the subjects who had successfully completed the learning phase had used similar strategies. All had at one point learned that they could classify 20 out of 24 faces correctly on the basis of face shape alone. They then had to learn how to discriminate the 2 thin-faced Harrys from the 10 thin-faced Charlies and the 2 fat-faced Charlies from the 10 fat-faced Harrys. Subjects mentioned the use of hair, mouth, eyes, and ears in making these discriminations. Since subjects' reports indicated that they were using similar strategies, a principal-components analysis was performed on the reaction time data to confirm this impression. Figure 6 shows the plot of the first two principal components from this analysis. The first principal component accounted for 49% of the variance; and the second, for 13.2%. Two groups of subjects could be

identified in the plot, but they were not as widely separated as had been the two groups in Experiment 2.

Group analyses were performed to determine if there were any interesting differences in the patterns of their reaction times. For Group 1, there was no significant difference between categories ($F \approx 1.0$) but a significant difference within categories, $F(22, 66) = 3.54$, $p < .01$. For this group, the comparison of the faces with characteristic face shape for their category to those with uncharacteristic face shape accounted for 32% of the variance between means, $F(1, 66) = 25.0$, $p < .01$. Within the Harry category, among those having the typical face, the comparison between faces with smile and those with frown was not significant. However, the comparison of those with hair to those without hair was significant, $F(1, 66) = 9.2$, $p < .01$, accounting for 12% of the variance. Within the Charlie category, among those having the thin face, the comparisons for smile–frown and for hair–no hair were not significant. However, there was a significant difference between the faces having ears and those without ears, $F(1, 66) = 11.0$, $p < .01$, accounting for 15% of the variance among the means.

From subjects' descriptions of their performance, it appeared that subjects in this group were using a sequential feature-testing approach to determine category membership. The first feature tested would be face shape. The second test would be for some other feature that allowed for the discrimination of the members of one category having a particular face shape from the members of the other category having the same face shape. For example, suppose that the face of a particular stimulus was found to be fat, then the subject would have to determine whether this face was one of the 10 fat-faced Harrys or one of the 2 fat-faced Charlies. Among the fat-faced Harrys, 6 had no hair and 4 had hair, whereas the 2 fat-faced Charlies had hair. Thus, a useful second feature to test for would be the presence of hair. If the head had no hair, then it could be classified as a Harry. If it did have hair, then further tests would have to be completed to determine whether

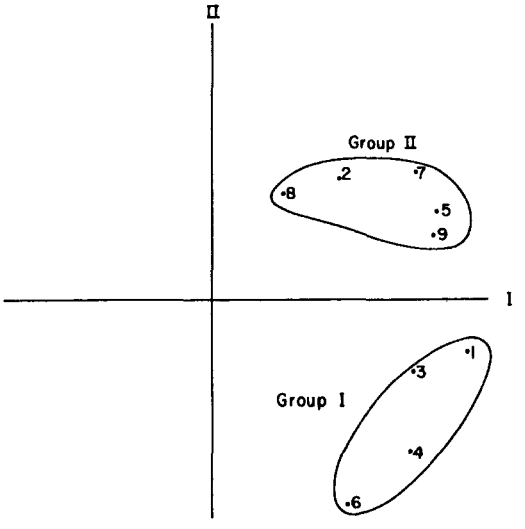

Figure 6. Inverse principal-components analysis of reaction times for subjects in Experiment 3.

it was a Harry or a Charlie. A hypothesized sequence of feature tests that would allow for the categorization of all the faces is shown in Figure 7. This sequence was developed based on subjects' reports and on the significant comparisons between means for different features. Mean reaction times for the set of faces that could be classified at each level in the hierarchy are shown on the graph. The comparison between the mean reaction time for Harry 3, which would be the last Harry to be classified, and Harrys 1, 2, and 10, which could be classified one test earlier in the sequence, accounted for 5% of the variance among the means but was not significant, $F(1, 66) = 3.86$, $.10 > p > .05$.

The comparison of the last Charlie to be classified, Charlie 2, to those classified one step previously, Charlies 1, 3, and 12, was significant, $F(1, 66) = 5.0$, $p < .05$, and accounted for 6.4% of the variance among the means.

For the second group of subjects, the analysis of variance indicated that there was a significant difference both between, $F(1, 4) = 15.3$, $p < .05$, and within, $F(22, 88) = 3.5$, $p < .01$, categories. Between categories, the mean reaction times were 884 msec for the Harrys and 958 msec for the Charlies. The comparison of reaction times for characteristic face shape to those for

odd face shape accounted for 71% of the variance. Within the Harry category, there was no significant difference for the comparison of those with smile to those with frown, for hair–no hair, or for ears–no ears. Within the Charlie category, the smile–frown and hair–no hair comparisons were not significant, but there was a significant difference between subjects' reaction times for those with ears and for those without ears, $F(1, 88) = 4.4$, $p < .05$, accounting for 5.8% of the variance among the means.

According to their subjective reports, these subjects took an approach somewhat opposite to that of the other group. Rather than learning which combination of features denoted a particular category, these subjects learned what features indicated that a face was not in a category. For example, one subject said, "A fat face was Harry, unless it had a frown and no ears, and hair; then it was a Charlie. A thin face was a Harry unless it was smiling and looking to the left." A plausible sequence of feature tests used by these subjects is shown in Figure 8. Again mean reaction times are shown for the faces that could be classified at a certain point in the hierarchy. According to this organization of tests, a fat face was a Harry unless it had both a frown

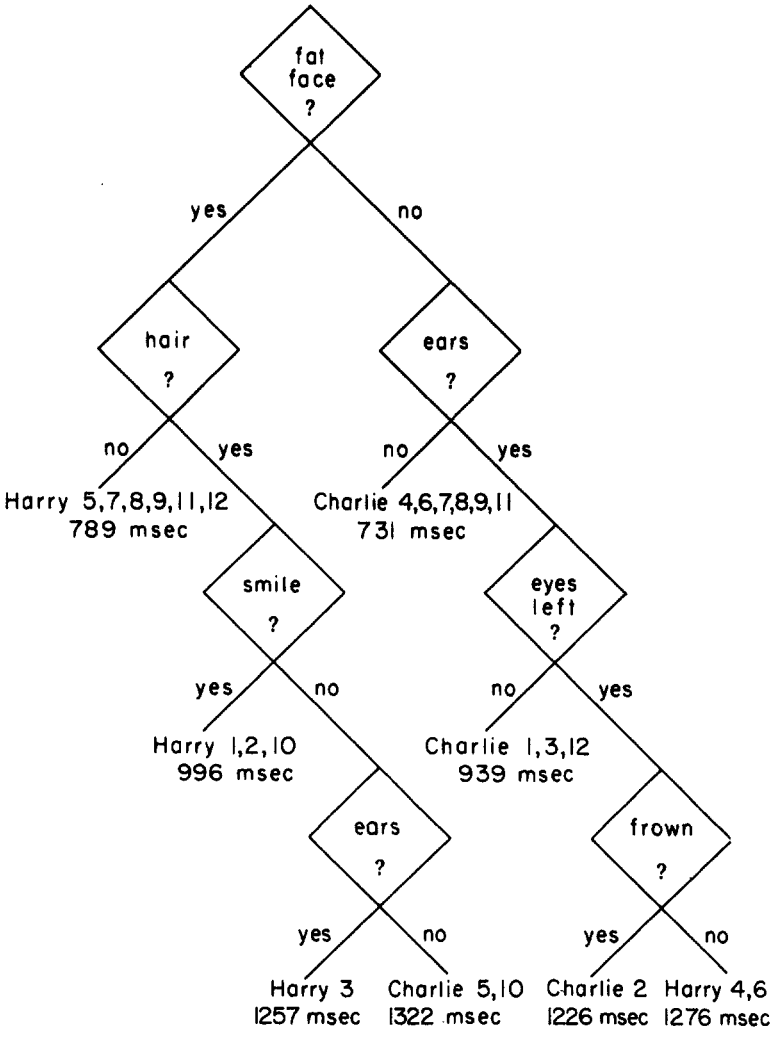

Figure 7. Sequence of feature tests for subjects in Group 1, Experiment 3.

and no ears. The only Harry having both of these characteristics was Harry 7, and Harry 7 had a significantly longer mean reaction time than the remaining Harrys, which had fat faces, $F(1, 88) = 4.0, p < .05$, accounting for 5.8% of the variance. On the other hand, a thin face was a Charlie unless it had a smile and ears. The only Charlies having these characteristics were Charlies 1 and 3. These had significantly longer reaction times than the Charlies without either of these features, $F(1, 88) = 4.7$, $p < .05$, accounting for 6.4% of the variance among the means.

An analysis of variance was performed on the typicality judgments for all subjects combined. Significant differences were found between the means of faces within each category. For the Harry category, 61% of the variance in the typicality judgments could be accounted for on the basis of face shape; and 26% of the variance, on the comparison of those with a smile to those having a frown. For the Charlie category, the same comparisons were significant, but with face shape accounting for 88% of the variance; and smile–frown, 11%. When analyzing the typicality judgments separately by the groups obtained in the reaction time analysis, the same comparisons were found to be significant for Group 2. For Group 1, in the Harry category, besides a significant difference for face shape and smile–frown, there was also a significant difference between the faces having ears and those without ears, $F(1, 33) = 5.04$, $p < .05$, accounting for 8.9% of the variance. In the Charlie category, for Group 1, face

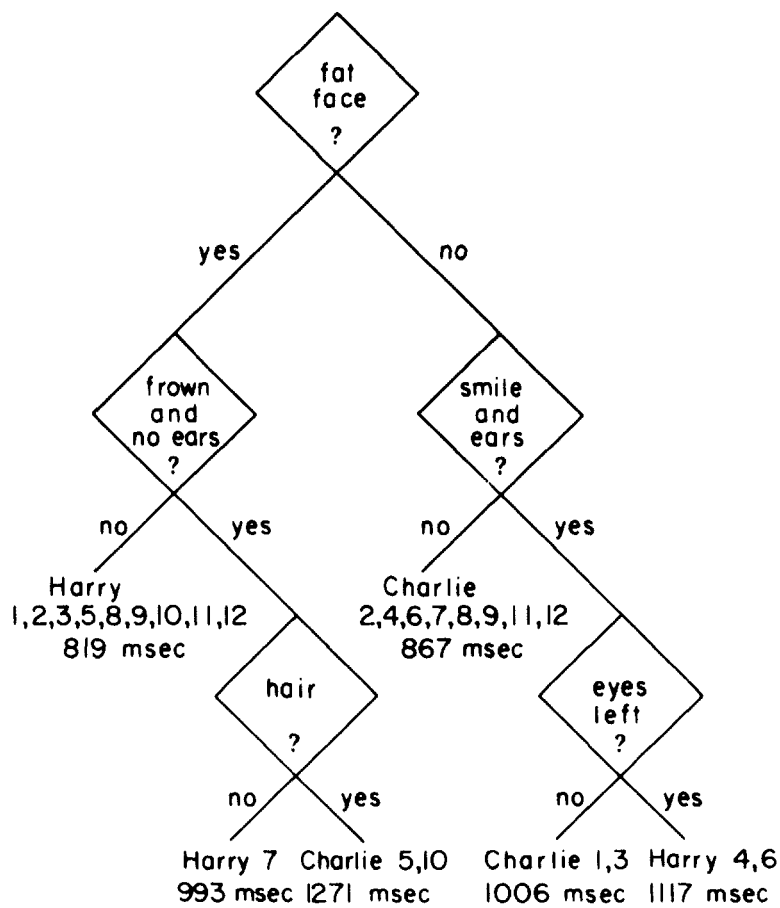

Figure 8. Sequence of feature tests for subjects in Group 2, Experiment 3.

shape was found to give rise to a significant comparison (accounting for 82% of the variance), but rather than the smile–frown comparison's being significant, the comparison between the faces having eyes looking to the left and those with eyes looking up was significant (accounting for 13% of the variance).

Similarity Judgments

An inverse principal-components analysis was performed on the uncentered Stimulus Pairs × Subjects matrix. The first principal component accounted for 60% of the variance, indicating a fairly high level of consistency across subjects. As in previous analyses, no distinct subgroups were found in the analysis of the centered matrix.

Thus, the groups of subjects found in the categorization task did not appear in the analysis of the similarity judgments. As in Experiment 2, although subjects used different strategies, the same features were employed in the tests. The order of the tests, however, did differ. Therefore, even if subjects had emphasized the features used in the categorization task in their similarity judgments, the same features would have been emphasized by both groups. Responses were therefore averaged for the multidimensional scaling analysis. The five-dimensional solution had a stress of .04; and the four-dimensional solution, a stress of .07. The five-dimensional solution was selected, since all five dimensions were interpretable. In the plane of the first two dimensions, shown in Figure 9, the faces segregated into four clumps on the basis of the possible combinations of smile–frown and fat or thin face. The third dimension corresponded to eyes looking to the left versus eyes looking up. The fourth dimension contrasted both hair versus no hair and nose shape. The fifth dimension was a

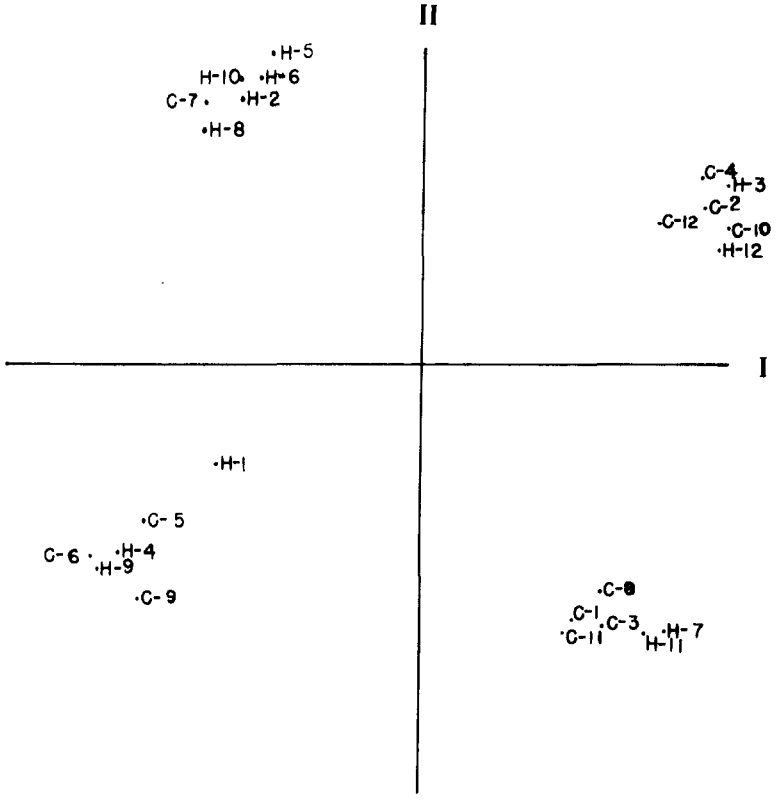

Figure 9. First two dimensions of multidimensional scaling solution for family resemblance condition after categorization. (C = Charlie; H = Harry.)

comparison of straight eyebrows to curved eyebrows.

Thus, as was found for the defining features stimuli, subjects were more analytical in their similarity judgments after having participated in the categorization task. Once again, a feature that was important in the categorization task (eyes) but was not a salient feature in the original scaling solution became an important feature in the similarity judgments after categorization. It is somewhat surprising that the contrast between ears and no ears did not turn up in this analysis, since both groups of subjects apparently used this feature in their strategies for classifying the stimuli.

Discussion

In comparing reaction times, typicality judgments, and rates of learning obtained in this study with the predictions of several models, it was found that the Rosch measure of family resemblance and the Medin and Schaffer context model made the best predictions for both categories. However, it is unlikely that these correlations are due to an overall degree of feature overlap between an exemplar and a category, as proposed by Rosch, or to the interactive similarity of exemplars, as proposed by the Medin and Schaffer model. Rather, it appears that a more basic process, a sequential testing of features, is operating. The correlations between the reaction times and the predictions of the Rosch and Medin and Schaffer models arise because subjects base their feature tests on features that allow them to classify the most stimuli, that is, on the features that are characteristic of one category, but not characteristic of the other. For example, face shape was highly informative of category membership; hence subjects incorporated a test for face shape in their categorization process. Since a particular face shape was highly characteristic of a category, faces having the characteristic face shape tended to have a higher feature overlap with other members of the category than faces with the uncharacteristic face shape; hence the correlation between reaction times and feature overlap. In the same manner, stimuli having the most informative features for category membership tend to be more similar to other members of their category and less similar to members of the contrasting category than stimuli that would be classified later in the sequence of feature tests, resulting in a correlation between the predictions of the Medin and Schaffer model and reaction times.

The lower correlations between the reaction times and the predictions of the Hyman and Frost models can be explained by noting that these predictions depended entirely on distances in the multidimensional space obtained in Experiment 1, and these distances were influenced mainly by smile–frown, face shape, and nose shape. According to the predictions of the Hyman and Frost models, faces in one category close to the boundary between the two categories in the plane of the first two dimensions would be predicted to have long reaction times even if these faces could have been classified quickly on the basis of features other than those represented in the scaling solution. Thus, the features of hair, ears, and eyes (features that subjects reported using to make their decisions about category membership) had no influence on distances in the multidimensional space. These features were given some weight in the predictions of the Rosch and Medin and Schaffer models.

As was found for the reaction times, correlations for the typicality judgments were highest for the Rosch and Medin and Schaffer models and lower for the Hyman and Frost models. These correlations result from subjects' basing their typicality judgments on features that were used in their feature tests. Since the analyses of variance on the typicalities for all subjects combined indicated that the comparisons for face shape and for smile–frown accounted for nearly all of the variance between mean typicality judgments, one might have expected higher correlations between typicalities and the predictions of the Hyman and Frost models. However, although smile–frown and face shape were the most important dimensions in the scaling solution, distance along the third dimension, that of nose shape, also affected the predictions of the Hyman and

Frost models. Since nose shape had no effect on typicality judgments, only moderate correlations were obtained between the Hyman and Frost models and the typicality judgments for the Harry category. The low correlations between these models and the typicalities for the Charlie category could be due to two factors. First, Charlie 4, which had an uncharacteristic face shape, and Charlie 10, which had an uncharacteristic mouth, were not positioned as close to the boundary between the two categories in the first two dimensions as were the other members of the Charlie category having these uncharacteristic features. Consequently these faces were predicted to have fairly fast reaction times and fairly high typicalities by the Hyman and Frost models, but, in fact, had long reaction times and low typicalities. Second, position of the eyes had a significant effect on the typicality judgments of Group 1 for the Charlie category; hence the averages for all subjects would also tend to reflect the influence of this feature. Position of the eyes had no effect on distance in the multidimensional space. Position of the eyes, in fact, had no effect on the predictions of the Medin and Schaffer and Rosch models either, since six members of each category had each type of eyes. Thus, the significant comparison among mean typicalities for eye position demonstrates the effect of the feature tests used by subjects in making their typicality judgments, even when the feature is neither characteristic of one category nor uncharacteristic of the contrasting category.

The results of the learning data reflect subjects' hypothesis testing. All subjects who met the criterion learned at one point that face shape alone could be used to classify 20 out of 24 of the faces. Hence, there was a significant effect for face shape on rate of learning. However, having made this discovery about face shape, subjects then had to learn to discriminate the odd-faced members of the category from the members of the other category having the same face shape. Some of the features used to make these discriminations were found to result in significant comparisons between the average trial of last error data (smile–frown for the Harrys and ears–no ears for the Charlies). Since the comparisons for face shape and smile–frown were found to be significant for the learning data for the Harry category, the correlations between average trial of last error and the predictions of the models were like those obtained for reaction times and typicality judgments. In the Charlie category, the presence or absence of ears had a greater effect on trial of last error than did face shape. Moreover, Charlie 2 had a very high trial of last error. Because none of the models gave a very high weight to the presence or absence of ears and because none predicted that Charlie 2 would be especially difficult to learn, low correlations were obtained for all models for the Charlie category for trial of last error. (The sequential feature tests proposed for Group 1 would, however, result in Charlie 2's being very difficult to learn.)

Although the sequential feature testing model and the models of Rosch and Medin and Schaffer often make the same predictions about the speed with which stimuli should be categorized, some discrepant predictions do arise when overall degree of feature overlap or interactive similarity is in opposition to the sequence of feature tests. For example, within the Harry category, Harry 1 had a family resemblance score of 41, and the evidence favoring its classification in the Harry category according to the Medin model was .960, whereas Harry 7 had a family resemblance score of 35, and the evidence favoring its classification was .828. However, within Subject Group 1, because of the order of feature tests employed, Harry 7 should have been categorized more quickly than Harry 1. Mean reaction time for Harry 7 for this group was 662 msec, whereas the mean reaction time for Harry 1 was 1,090 msec. Similarly, within the Charlie category, Charlie 2 had a family resemblance score of 39, and the evidence favoring its classification was .919, whereas Charlie 9 had a family resemblance score of 37, and the evidence favoring its classification was .852. However, according to the sequence of feature tests used by Group 1, Charlie 9 should have been categorized much more quickly than Charlie 2. Mean reaction time for Charlie 9 was 710 msec, and mean reaction time for Charlie 2 was 1,226 msec.

(For Group 2, it was impossible to come

up with examples that would differentiate between the feature testing model and the family resemblance models, since so many faces within a category were categorized at one level in the hierarchy. Moreover, the faces predicted to have the longest reaction times by the feature tests hypothesized for this group were also those having the lowest family resemblance scores and the least evidence favoring their classification within their own category.)

Further evidence favoring a sequential feature testing model is that, unexpectedly, higher correlations were obtained by the family resemblance and context models for the Harry category than for the Charlie category. In terms of the family resemblance or context models, eye direction should have been irrelevant to categorization, since half of the members of each category had each value of this feature. However, within both subject groups, a test for eye direction became useful at one point in the sequence of tests for determining members of the Charlie category. Thus, the lower correlations for the family resemblance and context models for the Charlie category would be expected. Even if one were to change the method for determining the parameters for the context model in order to maximize the correlations with reaction times, the correlations would still be lower for the Charlie category. There is no way to incorporate an effect for eye direction in the context model.

As in the experiment with the defining features categories, subjects' behavior was found to be like that of subjects in the concept formation studies. Subjects looked for rules that would enable them to classify the stimuli as easily as possible. Because the categories were structured so that no simple rule could serve to classify the stimuli (as was possible with the defining features categories), subjects were forced to develop complex sets of rules for determining category membership.

The fact that subjects could invent a sequence of rules for classifying these exemplars might be taken to indicate that these categories were not truly ill defined. However, this would be changing the meaning of *ill-defined* as it has been used in the categorization literature (Medin & Schaffer, 1978; Rosch, 1975). In the past, *well-defined* has been used to describe those categories that have features that are both necessary and sufficient for determining category membership. Categories that lack such necessary and sufficient features but instead have features that are only more or less characteristic of a category have been termed ill-defined. The category structure used in Experiment 3 fits this definition.

For ill-defined categories, it would *always* be possible to find some sequence of rules, though perhaps extremely complicated, that could be used for categorization. The question being addressed in Experiment 3 is, Given an ill-defined structure, how do subjects learn to classify the exemplars? The categories were designed so that no simple rule could serve to classify all the stimuli, and one might expect that a prototype or probabilistic process would be favored under these conditions. However, subjects apparently did develop rule systems. Considering the complexity of the rules shown in Figures 8 and 9, it should not be surprising that so many subjects were unable to learn the category membership of the faces within the allotted time. Other studies using ill-defined categories structured so that categories contained some of the characteristic features of the contrasting categories have also noted high failure rates for subjects (Medin & Schaffer, 1978; Rosch & Mervis, 1975). If the category concepts constructed by subjects were based on overall degree of feature overlap or interactive similarity to members of a category rather than on some analytic process such as the sequence of feature tests presented here, it is not obvious that one would expect the observed difficulty in learning the categories.

Conclusion

With the exception of the Smith et al. model, recent theories concerned with perceptual and conceptual categorization have assumed that even if exemplars can be decomposed into features, there is no logical rule or set of criterial features that is used to determine category membership. It has therefore been assumed that some probabilistic or holistic rather than analytical or logical process is used for determining

category membership. Various models that make use of the overall similarity or feature overlap of exemplars to category members or to category prototypes have been proposed. Even the Smith et al. model, which does assume the existence of defining features, nonetheless assumes a holistic first stage of processing to account for typicality effects on reaction times.

Experiments testing the various categorization models have been performed using artificial categories with a family resemblance structure, that is, with no set of criterial features. Because no simple rule could be used for classifying all the stimuli in these experiments, it was assumed that the findings of the concept formation studies were irrelevant to these stimulus structures and that subjects would not look for logical rules. The results of both the categorization experiments presented here indicate that whether or not categories have defining features, subjects attempt to develop rules for classifying the stimuli into categories. Although no simple rules were possible for the family resemblance structures, this did not deter subjects from using this approach, even though very complex rule systems were necessary. The complexity of the necessary rules resulted in a high percentage of subjects' failing to learn the categories.

In developing their rule system, subjects attempt to find the most efficient set of rules (and hence the smallest number of rules to remember). In a defining features case, the presence of the combination of defining features in a category or their absence in the contrasting category will be the simplest rules that could be used to classify the stimuli. (Subjects in Experiment 2 used both of these strategies.) In a family resemblance case, the rules allowing for the classification of most faces would be tests for the presence of the most characteristic features of a category that were not also characteristic of the contrasting category (e.g., face shape in Experiment 3). Thus, faces having a high degree of family resemblance or similarity to their own category and a low degree of family resemblance or similarity to the contrasting category would tend to be classified fastest. Hence, the predictions of all the family resemblance models would tend to be correlated with the obtained reaction times.

A relation between these findings and the work of Garner (1974) and his associates should be noted. Whitman and Garner (1963) studied the effect of category structure on ease of learning. Unlike the categories in the concept formation studies of Bruner et al. (1956) and Neisser and Weene (1962), the categories used by Whitman and Garner did not contain all possible combinations of feature values. Whenever less than the total set is used (as was the case in our experiments), a correlational structure between features is introduced. Whitman and Garner differentiated two types of correlational structure, simple and complex. In a simple structure, two or more dimensions are perfectly correlated; in a complex correlational structure, correlations exist only on the basis of the interactions of features. They found that categories having a simple structure were much easier to learn than those having a complex structure. In the defining features condition of the present study (Experiment 2), the stimuli came close to having a simple structure. In the Harry category, round nose always occurred with hair, whereas in the Charlie category this combination never occurred. In the Charlie category, eyes looking upward always occurred with frown. For the family resemblance condition, more complex correlations existed. For example, in the Harry category 6 stimuli had fat faces and ears, but none of the faces in the Charlie category had this combination of features. In line with Garner's findings, we found that the defining features condition was much easier to learn than the family resemblance condition. However, although Garner (1974) chose to interpret these kinds of results by focusing on stimulus structure rather than on processing, we have stressed that the ease or difficulty of learning results from subjects' attempts to use the same type of process for both simple and complex structures.

In most research in information processing, it is assumed that subjects behave similarly and that therefore one can average

data across subjects. However, a few studies have shown the importance of examining individual data (e.g., Cooper & Shepard, 1973; Hock, Gordon, & Corcoran, 1976). In the Cooper and Shepard study and in that of Hock, Gordon, and Corcoran, it was found that different subjects used different types of processes in performing the same task. In the present research subjects were found to use the same type of process (that is, they developed rules for classification) but to differ in the particular rules used. Thus, in categorization tasks, considering only total group averages summed across individuals who may have used different rules can lead to incorrect conclusions about the types of processes that underly categorization. For example, suppose that a stimulus set had four features —A, B, C, and D, with A being the most characteristic feature of Category 1. If one group of subjects used a rule based on the combination of Features A and B, another group a rule based on A and C, and another group a rule based on A and D, the data averaged across all subjects would appear to show that all features had an effect on reaction time as predicted by a feature overlap, prototype, or exemplar model. However, when looking at individual groups of subjects, reaction times might only show an effect of those features used by the particular group in making their categorizations. Chumbley, Sala, and Bourne (1978) have made this point in their categorization study in which analyses of individual data revealed large individual differences in the features used by subjects in making category judgments.

Our experiments were concerned primarily with studying the type of processing used in categorization, given that categories have a particular structure, and not directly with studying category structure. However, considered together with other relevant findings, these data are relevant to the issue of category structure. In Experiment 2, we found that one group of subjects discovered the features that defined one of the categories, and a second group discovered the complement of these features in the other category. For the group that was aware of the defining features, no significant differences in reaction times were obtained among the members of a category. This finding could be taken to suggest that if subjects were aware of defining features in natural categories one would not expect within-category differences in reaction time. However, within-category differences in classification time for members of natural categories is a well-established finding (Caramazza et al., 1976; Rips et al., 1973; Rosch, 1975). Thus, one might conclude that natural categories are not defined by a conjunction of criterial features. In the family resemblance experiment (Experiment 3), within-category differences in reaction times were obtained. However, it would be incorrect to conclude from this that internal category representation must be ill-defined in the sense of consisting of a loose collection of features or being organized around a prototype. Subjects looked for a set of deterministic rules for classification in Experiment 3 rather than using a process based on a global measure of family resemblance, such as feature overlap or distance to a prototype. It may be that subjects' internal representation of the category directly reflects a decision tree structure similar to that in Figure 8.

On the other hand, one cannot rule out the possibility that within-category differences in reaction time could be obtained for artificial categories having criterial features if feature values were other than discrete and well defined. For example, given stimuli that had features that varied continuously, a defining feature might be a long nose for one category and a short nose for the other. Reaction time for deciding whether a face had a long or short nose would depend on how long or short it was, with intermediate values resulting in longer reaction times.

A similar argument could be made for categories having features that were ill-defined in some respect other than taking on a range of values along a specific continuum. For example, a defining feature of the category *chair* might be that a chair has a back. However, not all chairs have backs to the same degree. (Consider, for example, a kitchen chair versus a bean bag chair.) Differences in reaction time for

classification among members of a category would be expected if the members possessed the defining features to differing degrees. The time it takes to verify whether a feature is present may be related to the degree to which a part of a stimulus could be construed as satisfying the internally represented value for that feature (Brownell & Caramazza, 1978). Differential predictions between such a model based on criterial features and a prototype or probabilistic model could be worked out on the basis of the effects of nondefining features on categorization time (Caramazza & Brownell, Note 2).

As stated in the introduction, we have assumed that subjects use a featural analysis as part of the categorization process. The type of stimuli used in the experiments we have reported certainly allows subjects to identify and use individual features. However, even with stimuli less obviously composed of features (i.e., dot configurations), Barresi et al. (1975) found that subjects searched for distinctive features in their attempts to determine category membership. The fact that Posner and Keele (1968) and many others have found evidence consistent with a prototype model for holistic stimuli does not preclude the possibility that subjects could have been using a sequence of feature tests for these stimuli as well. As has been pointed out many times before, prototypes are also those stimuli possessing the features most common to members of a category (Barresi et al, 1975; Neumann, 1977). Thus, correlations between reaction times and similarity to a prototype would be expected on the same basis as correlations with family resemblance, even if subjects were using a sequence of feature tests to assign category membership.

Finally, we do not want to overstate the relevance of the results we have reported to deciding issues of category representation and categorization, especially by minimizing potential problems in generalizability to other stimulus domains. However, it would be equally unacceptable to ignore the possibility that the sequential feature testing model could plausibly be extended to cover categorization of artificial categories with continuous or less well-defined features. Generalization to natural categories is, of course, more difficult; but this is a difficulty that presently remains undefined. The important conclusion to draw from the present experiments is that previous findings with artificial categories, which have been used to support probabilistic or similarity models of categorization (Medin & Schaffer, 1978; Reed, 1972; Rosch et al., 1975), do not necessarily support such models but may, in fact, result from subjects' use of a logical sequence of rules for categorization.

Reference Notes

1. Kruskal, J. S., Young, F. W., & Seery, J. B. *How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding.* Bell Laboratories Technical Report, Murray Hill, N.J., 1977.
2. Caramazza, A., & Brownell, H. *Categorical effects in perceptual judgments and classification of objects.* Unpublished manuscript, Johns Hopkins University, January 1979.

References

Barresi, J., Robbins, D., & Shain, K. Role of distinctive features in the abstraction of related concepts. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, *104*, 360–368.

Bierwisch, M. Semantics, In I. J. Lyons (Ed.), *New horizons in linguistics*. Baltimore, Md.: Penguin Books, 1970.

Bolinger, D. L. The atomization of meaning. *Language*, 1965, *41*, 555–573.

Bourne, L. E. Knowing and using concepts. *Psychological Review*, 1970, *77*, 546–556.

Bourne, L. E., Ekstrand, B. R., & Dominowski, R. L. *The psychology of thinking*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Brownell, H. & Caramazza, A. Categorizing with overlapping categories. *Memory & Cognition*, 1978, *6*, 481–490.

Bruner, J. On perceptual readiness. *Psychological Review*, 1957, *64*, 123–152.

Bruner, J. S., Goodnow, J. J., Austin, G. A. *A study of thinking*. New York: Wiley, 1956.

Caramazza, A., Hersh, H., & Torgerson, W. Subjective structures and operations in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 103–117.

Chumbley, J., Sala, L. S., Bourne, L. Bases of acceptability ratings in quasinaturalistic concept tasks. *Memory & Cognition*, 1978, *6*, 217–226.

Clark, E. V. What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press, 1973.

Cooper, L. A., & Shepard, R. N. Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press, 1973.

Fillenbaum, S., & Rapoport, A. *Structures in the subjective lexicon*. New York: Academic Press, 1971.

Garner, W. *The processing of information and structure*. Potomac, Md.: Erlbaum, 1974.

Goldman, D., & Homa, D. Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 375–385.

Hayes-Roth, B., & Hayes-Roth, F. Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, 1977, *16*, 321–338.

Henley, N. A. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, 1969, *8*, 176–184.

Hock, H. S., Gordon, G. P., & Corcoran, S. K. Alternative processes in the identification of familiar pictures. *Memory & Cognition*, 1976, *4*, 265–271.

Homa, D., & Chambliss, D. The relative contributions of common and distinctive information on the abstraction from ill-defined categories. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, *104*, 351–359.

Hutchinson, J. W., & Lockhead, G. R. Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 660–678.

Hyman, R., & Frost, N. H. Gradients and schema in pattern recognition. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance*. London: Academic Press, 1975.

Katz, J. J. *Semantic theory*, New York: Harper & Row, 1972.

Katz, J. J., & Fodor, J. A. The structure of a semantic theory. *Language*, 1963, *39*, 170–210.

Kruskal, J. D. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 1964, *29*, 1–27.

Leech, G. *Semantics*. Baltimore, Md.: Penguin Books, 1974.

McCloskey, M., & Glucksberg, S. Decision processes in verifying class inclusion statements: Implication for models of semantic memory. *Cognitive Psychology*, 1979, *11*, 1–37.

Medin, D. L., & Schaffer, M. M. A context theory of classification learning. *Psychology Review*, 1978, *85*, 207–238.

Miller, G. A., & Johnson-Laird, P. N. *Language and perception*. Cambridge, Mass.: Belknap Press, 1976.

Neisser, U., & Weene, P. Hierarchies in concept attainment. *Journal of Experimental Psychology*, 1962, *64*, 640–645.

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

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

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

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

Rips, L. J., Shoben, E. J., & Smith, E. E. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 1–20.

Romney, A. K., & D'Andrade, R. G. Cognitive aspects of English kin terms. In A. K. Romney & R. G. D'Andrade (Eds.), Transcultural studies in cognition. *American Anthropologist*, 1964, *66*(3, Pt. 2), 146–170.

Romney, A. K., Shepard, R. N., & Nerlove, S. B. *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 2). New York: Academic Press, 1972.

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

Rosch, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 1975, *104*, 192–233.

Rosch, E., & Mervis, C. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 1975, *7*, 573–605.

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

Smith, E. E., Shoben, E. J., & Rips, L. J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 1974, *81*, 214–241.

Tucker, L. R., & Messick, S. An individual differences model for multidimensional scaling. *Psychometrika*, 1963, *28*, 333–367.

Whitman, J. R., & Garner, W. C. Concept learning as a function of form of internal structure. *Journal of Verbal Learning and Verbal Behavior*, 1963, *2*, 195–202.

Wittgenstein, L. *Philosophical investigations*. New York: MacMillan, 1953.

Received February 20, 1979 ■