Memory & Cognition
1978, Vol. 6 (3), 217-226

Bases of acceptability ratings in quasinaturalistic concept tasks

JAMES I. CHUMBLEY
University of Massachusetts, Amherst, Massachusetts 01002

and

LINDA S. SALA and LYLE E. BOURNE, JR.
University of Colorado, Boulder, Colorado 80302

The extent to which subjects used a probabilistic or a distance judgment process in rating the acceptability of concept exemplars was studied. After viewing 100 descriptions of airline uniforms chosen as fashionable by the public, subjects rated all possible uniforms for their public acceptability. The uniforms were described in terms of four three-valued qualitative (e.g., type of fabric) or quantitative (e.g., jacket length) attributes, with 20 subjects assigned to each description type. The results indicated that the majority of subjects in both groups used some form of probabilistic decision rule, with very few, if any, using a rule based on distance from a prototype. A data analysis technique which permitted examination of the behavior of individual subjects documented a significant amount of inter- and intrasubject variability in the number of aspects incorporated into the judgment process, in the number of different judgment rules utilized for different aspects, and in the types of decision rules used. It was concluded that subjects in earlier studies might well have been attending primarily to frequency differentials but that various methodological difficulties with the acquisition procedures and with the level of data analysis produced ambiguous and sometimes contradictory results.

Until recently, most studies investigating conceptual behavior utilized deterministic or logical class concepts. These concepts permit inference of the conceptual class of an object, given knowledge of the relevant attributes and the rule or relation (e.g., conjunction) to be used in partitioning the set of stimulus objects (Bourne, 1974). In contrast, many recent experiments used probabilistic concepts, sometimes referred to as natural or semantic concepts because of their similarity to real-world categories. With a probabilistic concept, the subject cannot determine with certainty the "correct" conceptual class of any stimulus object (cf. Reed, 1972). Since with deterministic concepts the rule and relevant attributes were well defined, the major interest of researchers was in the way subjects identified the relevant attributes (Levine, 1975; Millward & Wickens, 1974) or in the way subjects learned the rule relating the relevant attributes to the response classes (Bourne, 1974). While some researchers have exhibited similar interests in studies using probabilistic concepts (e.g., Dansereau & Brown, 1974; Homa & Vosburgh, 1976; Keele, 1973; Peterson, Meagher, Chait, & Gillie, 1973; Posner, 1973), others have asked a question which did not arise in earlier studies of deterministic class concepts, namely, what is the nature of the decision rule by which subjects classify stimulus objects? The experiments that addressed this question (Barresi, Robbins, & Shain, 1975; Beach, 1964a; Hyman & Frost, 1975; Neumann, 1977; Reed, 1972; Reed & Friedman, 1973; Rosch & Mervis, 1975) have not been able to answer it with any degree of precision and, in many cases, the answers are contradictory. A similar state of affairs exists in recognition memory studies where conceptual behavior is involved (e.g., Bransford & Franks, 1971; Franks & Bransford, 1971; Neumann, 1974; Reitman & Bower, 1973; Hayes-Roth & Hayes-Roth, Note 1).

Thus, to date, we appear to have no better statement of what is learned than that of Attneave (1957), who considered it likely that the subject "learns something about at least three characteristics of the class: (a) its central tendency; (b) *how* its members may differ from one another, i.e., in what properties, or on what dimensions; and (c) its dispersion, i.e., *how much* its members may differ from one another on the several dimensions of variability" (p. 87).

Although one of the difficulties in determining precisely what subjects learn in a probabilistic concept

This research was conducted while the first author was on sabbatical leave at the Institute for the Study of Intellectual Behavior at the University of Colorado and is Publication 72 of the Institute. Support for the research was provided by a U.S. Public Health Service special fellowship to the first author, by Research Grant MH-14314 and Research Scientist Award 1-K5-MH-37497 from the National Institute of Mental Health, and by Research Grant GB-340-77X from the National Science Foundation.

task may be attributed to the large number of closely related decision strategies that are possible (Reed, 1973), there are other problems of a more methodological nature. First, it is not clear how much of the variance in answers to the question of what is learned can be attributed to differences in stimulus materials and tasks. Stimuli have varied from random dot patterns (Barresi et al., 1975; Hyman & Frost, 1975), to polygons, tone sequences, and columnar patterns generated from Markov rules (Aiken & Griffin, 1972; Dansereau & Brown, 1974), to stimuli with well-defined quantitative dimensions (Neumann, 1977; Reed, 1972; Reed & Friedman, 1973), and to stimuli with well-defined qualitative dimensions (Beach, 1964a; Neumann, 1974; Hayes-Roth & Hayes-Roth, Note 1). Studies in which concepts varied along quantitative stimulus dimensions have usually produced results supporting a decision strategy incorporating a measure of distance from a prototype, whereas the results of studies using qualitative stimulus dimensions tend to support decision models based on likelihood judgments derived directly from the probability distribution of stimulus values. Studies using random dot patterns have produced mixed results.

A second source of difficulty in assessing the results of research on probabilistic concepts is the significant intersubject variability in how stimulus information is used in making classifications (Hyman & Frost, 1975; Reed, 1972; Reed & Friedman, 1973). There is evidence that, in generating a response, subjects vary in the degree to which they attempt to utilize all of the available stimulus information (Beach, 1964a). Unfortunately, the methods that have been used to determine the strategy used by an individual subject have been quite imprecise, varying from verbal reports to correlational evidence. In the case of the correlational evidence, the predictions of different decision strategies are highly correlated, making discrimination of strategies very difficult (Hyman & Frost, 1975).

The research reported here addresses the question of what is learned in a way which minimizes the above difficulties. First, the question of the effect of type of dimension, qualitative or quantitative (numeric), is explored directly. Second, a new methodology is introduced which permits both a direct assessment of the behavior of an individual subject and an assessment of each subject's behavior with respect to each stimulus dimension. Finally, the characteristics of the stimulus distributions used for acquisition simplify the predictions of the main classes of decision models, making a discrimination between them less difficult.

Reed (1972, 1973) proposed that there are two general classes of decision models, probability models and distance models. Probability models utilize decision rules similar to those found in nonparametric statistical techniques. The value of a stimulus attribute has only nominal relevance and is not used in the decision function. Only the relative frequency of occurrence of each of the dimension values is incorporated into the decision function. Distance models use a logic that is more similar to parametric statistical techniques in that the value of each stimulus attribute is incorporated into the decision function along with the frequency of its occurrence. Examples of probability models are the cue-validity model of Beach (1964b), the attribute frequency model of Neumann (1974), the schematic model of Hayes-Roth and Hayes-Roth (Note 1), and the family resemblance model of Rosch and Mervis (1975). Examples of distance models include the proximity, average distance, and prototype models presented by Reed (1972, 1973) and the rule model presented by Hyman and Frost (1975). The experiment reported below provides information about the relative incidence of usage of these two classes of decision models and explores the effect of type of dimension value, qualitative or numeric, on the usage frequency.

Table 1 displays the information about stimulus dimensions that is pertinent to distinguishing between classes of models. In the experiment, subjects were shown a sample of 100 uniforms that were members of a conceptual class, "attractive" uniforms, and then asked to rate some new and some old uniforms with respect to the degree to which each one belonged to the class. The stimuli had four dimensions. For subjects in the qualitative-dimension group, the dimensions were color, fabric, pants type, and sweater type. For subjects in the numeric-dimension group, they were jacket length, boot height, number of buttons, and number of pockets. Recall that probability models predict that subjects should be sensitive only to the relative frequency of occurrence of dimension values and should tend to rate stimuli as a function of the relative frequency of the dimension values displayed. For example, a stimulus with four high-frequency (HF) values should receive a higher rating than a stimulus with four medium-frequency (MF) or four low-frequency (LF) values. In addition, one would expect subjects to adopt this decision strategy irrespective of dimension type. That is, relative frequencies can be computed for both quantitative and qualitative dimensions.

The prototype-distance model, which Reed (1972) and Reed and Friedman (1973) concluded their subjects were using, makes quite different predictions. The prototype model assumes that subjects calculate a measure of central tendency (usually the mean) for each stimulus dimension frequency distribution. The prototype for a conceptual class is an abstract entity with dimension values equal to the means for each dimension. When asked to rate the degree to which a particular stimulus fits a category, subjects determine the distance of the test stimulus from the prototype. Stimuli that are close to the prototype would be given high ratings; stimuli that are more distant would be given low ratings. Since the mean value on each numeric-valued stimulus

Table 1
Stimulus Dimensions, Their Values, and the Frequency of the Values

Relative Acquisition Frequency	Acquisition Frequency	Dimension							
		Color	Fabric	Pants	Sweater	Jacket Length (in.)	Boot Height (in.)	Number of Buttons	Number of Pockets
High	60	Blue	Mohair	Flair	V Neck	33	9	4	3
Medium	30	Orange	Wool	Straight	Crew Neck	32	10	5	2
Low	10	Gray	Nylon	Leotards	Turtleneck	31	11	6	1
Mean						32.5	9.5	4.5	2.5

dimension in Table 1 is midway between the HF and MF values, the stimulus with four MF values is just as close to the prototype as the stimulus with four HF values and should, therefore, receive the same rating. Since LF values are further from the dimension mean than HF and MF values, the stimulus with four LF values should receive a much lower rating. In addition, one would not expect a distance model to be used as a basis for responding when the stimulus dimensions are qualitative and truly discontinuous since calculation of a mean (or median) would be logically untenable.

The distributions of stimulus values and the covariation of values across dimensions were selected to simplify predictions from the models. First, many more stimuli were used during the acquisition phase of the experiment than in most previous experiments, Beach's (1964a) experiment being a significant exception. Use of a large number of stimuli permitted several desirable controls. As can be seen in Table 1, the distributions were deliberately skewed so that the prediction that subjects would rate most highly the dimension values close to the central tendency would not be confounded with any tendency of subjects to rate a middle scale value most highly. The large number of acquisition stimuli allowed 10 presentations of the LF values even with the skewed distributions. In addition, use of a large number of acquisition stimuli made it possible for stimulus dimensions to be pairwise independent. Some previous studies had rather unusual conjoint frequencies of dimension values. Nonindependence of stimulus dimensions is obviously a common, but not universal, occurrence in the natural world, but its presence complicates discriminating between decision models. An example of the difficulties that can arise can be found in the Reed (1972) and Reed and Friedman (1973) experiments, in which the combination of a small number of acquisition stimuli, skewed distributions, and dimensional dependence leads to the prediction that some test stimuli, for example, a person 30 years of age making $12,000 per year or a person 40 years old with one child, would not belong to either conceptual class.

A second property of the stimulus dimension distributions in Table 1 is that, for the numeric dimensions, there is homogeneity of variance. Since distance is a function of the variance of a distribution, unequal variances produce unequal effects of dimensions on distance from the prototype unless subjects use something like a z transform in calculating distances. It is not at all clear how subjects adjust for unequal variances, so it was decided that the most reasonable course of action was to use equal variance distributions.

Reed (1972) and Reed and Friedman (1973) found that the fit of the prototype model to their data could be improved by weighting the stimulus dimensions by their objective validity for discriminating between conceptual categories. They also found, however, that subjects differed in their sensitivity to objective validity differences and in the degree to which they applied validity standards derived from extra-experimental experience (see Reed and Friedman, 1973, p. 160, for examples). We attempted to minimize ambiguities in interpretation of the results due to dimension differences in objective and subjective validity or importance. Dimension-value distributions were identical across distributions so that the objective validity or importance was the same for all dimensions. To minimize subject-related differences in dimension validity, the dimensions and their values were chosen so that, in our opinion, there would be no strong extra-experimental bias.

One final aspect of the experimental design should be noted. Interspersed among groups of concept exemplars presented during acquisition were blocks of test trials with feedback. The test trials provided subjects with practice in using the rating scale. For the test trials with feedback, stimuli were chosen such that the most viable probability and distance models would predict the same rating. In effect, the test trials with feedback only told the subject that the more HF values displayed in a stimulus, the higher should be its rating. None of the stimuli contained MF values, the values about which the models make differential predictions. Another function of the test trials with feedback is that they provide both a means of assessing the degree to which each subject is discriminating among stimuli immediately after acquisition and a means for determining the degree to which forgetting takes place during the final test series.

METHOD

Subjects

Students in introductory psychology classes at the University of Colorado elected to participate in the experiment as partial fulfillment of a class requirement. The experimental design

called for two groups of 20 subjects each, but a total of 44 students participated since the data for four subjects were discarded because of failure to follow instructions. Of these subjects, three indicated on a postexperimental questionnaire that they had responded either randomly or only on the basis of a preexperimental personal preference. One subject gave the same response for all test stimuli.

Procedure

The concept learning task was presented in the context of a cover story. The story asked the subject to imagine being a member of a futuristic society who was applying for a job as a marketing analyst for an airline. According to the story, the airline was designing new unisex uniforms for its workers. The subject was to discover what kinds of uniforms the public found appealing by viewing, on a display screen, descriptions of uniforms selected as attractive by a sample of the public. During test trials, the subject rated each uniform description on a scale from 1 to 5, with 1 indicating a low public rating and 5 indicating a high public rating. The instructions emphasized that the subject's personal preferences were irrelevant, since the corporation was interested only in the public opinion and the subject's ability to gauge it.

Acquisition. During the acquisition phase of the experiment, subjects were presented with uniform descriptions corresponding to the choices of the sample of 100 members of the public. As explained below, only 44 unique descriptions were used as sample stimuli during acquisition. Each sample stimulus was presented for 3 sec, with no response required. Interspersed with the sample stimuli were 34 practice test trials with feedback. On these test trials, the subject was given as much time as required to respond. Following the response on the feedback trials, the correct response was superimposed for 2 sec on the display screen below the uniform description just rated.

The stimulus presentation order was completely random with the exception of restrictions on the blocking of the types of trials, sample stimulus, or test with feedback. Before each change in type of trial, the subject was informed of the transition. Thus, it was very clear to subjects which stimuli should be used in forming a concept of an attractive uniform and which were being presented to test the adequacy of their concept and familiarize them with the rating scale. There were six unequal-size blocks of trials: 40 presentations of sample stimuli, 9 test trials with feedback, 30 more sample presentations, 9 more tests with feedback, the remaining 30 exemplars, and a final block of 16 tests with feedback.

Test. Following acquisition, subjects were tested in three different ways, two using the display terminal. First, each subject rated, without feedback, the entire set of 81 possible uniform descriptions. The presentation order was completely random and responding was self-paced. After these tests, subjects were presented with the three values of each stimulus dimension, one dimension at a time. For each dimension, they were asked to indicate the dimension value they thought to be ideal. Third, a postexperimental questionnaire also asked subjects to rank order the dimension values on each dimension.

Stimulus Materials

The stimuli were represented to the subjects as verbal descriptions of some attributes of uniforms. Subjects in the qualitative-dimension group saw descriptions using the dimensions of color, fabric, pants style, and sweater style. Numeric-dimension subjects had descriptions in terms of jacket length, boot height, number of buttons, and number of pockets. A stimulus was a set of four dimension labels horizontally arrayed across a viewing screen with a dimension value listed below each label. The values for the dimensions used are displayed in Table 1. Since each of the four dimensions had three values, 81 unique stimuli were possible for each stimulus set.

Skewed frequency distributions of the dimension values were used to define the concept for the numeric dimensions. For each dimension, the direction of skew was determined randomly and the acquisition frequencies of the dimensions values were assigned appropriately. For the qualitative dimensions, the frequency-value pairings were randomly made, since there was no natural ordering of values. The resulting frequency-value pairings may be seen in Table 1.

Of the 81 possible stimuli, only 44 unique stimuli were used in the acquisition phase of the experiment. The individual stimulus frequencies in the set of 100 stimulus presentations in acquisition ranged from 1 to 13. The stimuli used during acquisition were chosen to meet two restrictions. First, the marginal frequencies of stimulus dimension values were equal to those shown in Table 1. Second, all joint frequencies of pairs of dimension values were equated at their expected frequencies for all dimension pairs. For, example, the joint frequency of a HF value on one dimension with a MF value on another dimension was 18, $.6 \times .3 \times 100$.

A total of 16 stimuli were used for the test trials with feedback.These stimuli can be grouped into five classes: a stimulus with four HF values, four stimuli with three HF and one LF values, six stimuli with two HF and two LF values, four stimuli with one HF and three LF values, and a stimulus with four LF values. The five classes will be designated by the number of HF values (zero to four) present in the stimulus. All 16 stimuli were used in the final block of 16 test trials with feedback. In the first two blocks of nine test trials with feedback, the 4HF, the 0HF, two of the 3HF, two of the 1HF, and three of the 2HF stimuli were used.

Apparatus

The experiment was conducted using the CLIPR real-time computing facility at the University of Colorado. Stimulus sequencing and presentation, timing control, and data collection were under computer control. Up to six subjects participated simultaneously at independent stations. Each subject was presented with instructions and stimuli on a CRT display terminal and responded using a set of response buttons. Subjects were run asynchronously and each subject had a different random order of stimuli.

RESULTS

The basic data are the ratings of the 81 uniform descriptions by each subject on the final block of test trials with feedback and during the test phase.

Analyses of Group Data

Data from the test trials with feedback. For stimuli containing only HF and LF values, the most commonly discussed decision models assign ratings that are a simple linear function of the number of HF values, viz., the response should equal the number of HF values plus one. After subjects had completed viewing the 100 sample presentations and practiced responding for two blocks of nine test trials with feedback, they were tested on 16 stimuli in the last block of test trials with feedback. In general, subjects were not very accurate in making exactly the response predicted by the models. They were exactly correct on only 55% of the trials. However, each subject's average responses corresponded very well (with some regression toward the mean) to the correct model-predicted responses for each of the five stimulus types defined by the number of HF values. Table 2 presents the mean response to each stimulus class and the proportion correct on the class both for the last block

Table 2
Stimulus Ratings (R) and Proportion of Correct Responses (P) For Stimuli Used on Test Trials With Feedback

Stimulus Class	Correct Response	Qualitative Dimensions: Final Block of Test Trials With Feedback P	Qualitative Dimensions: Final Block of Test Trials With Feedback R	Qualitative Dimensions: Final Test Series P	Qualitative Dimensions: Final Test Series R	Numeric Dimensions: Final Block of Test Trials With Feedback P	Numeric Dimensions: Final Block of Test Trials With Feedback R	Numeric Dimensions: Final Test Series P	Numeric Dimensions: Final Test Series R	Average P	Average R
4HF	5	.75	4.60	.75	4.60	.65	4.45	.65	4.30	.700	4.487
3HF	4	.58	3.66	.55	3.86	.52	3.75	.49	3.68	.534	3.737
2HF	3	.59	3.06	.59	3.07	.42	3.06	.46	2.88	.517	3.015
1HF	2	.44	2.49	.55	2.36	.50	2.45	.49	2.24	.494	2.384
0HF	1	.55	1.65	.60	1.75	.35	1.75	.50	1.80	.500	1.737

of 16 test trials with feedback and for the same stimuli when tested during the final series of 81 trials in the test phase. Two analyses of variance, one on average rating and the other on proportion correct data, indicated that the only effective variable was the number of HF values in the stimulus. For the average ratings, all stimulus classes differed from each other [$F(4,152) = 138.18$, $p < .001$, standard error of the difference between means (SED) = .131]. For proportion correct, the 4HF stimulus differed from the other classes, producing an overall significant difference [$F(4,152) = 4.74$, $p < .005$, SED = .056]. Type of stimulus dimensions (qualitative vs. numeric) and time of rating had essentially no effect on either average stimulus rating or proportion correct ($p > .25$ in all cases). Finally, additional analyses of the data from these 16 stimuli indicated that, for all four stimulus dimensions for each of the groups, uniform descriptions containing the HF value of a dimension were rated significantly higher than descriptions containing the LF value [smallest $F(1,19) = 11.08$ $p < .005$]. This means that each dimension was being used by at least some subjects in generating rating responses.

Data from the test phase. Separate within-subject analyses of variance were conducted for the qualitative- and numeric-dimensions groups. Each factorial analysis had four factors, the four stimulus dimensions, each with three values. Thus, a rating response of a uniform description is assumed to be some function of the responses to each of the four dimensions, and significant variability associated with a dimension implies some subjects are attending to the dimension. The analyses of the average ratings of the 81 test trial stimuli (including the 16 stimuli which had also been used on test trials with feedback) varied as a function of the dimension values displayed for all four dimensions for each group (all $ps < .005$). Table 3 presents the mean ratings, Fs, and SEDs for the test trial data. Note that for all dimensions, descriptions displaying HF values produced significantly (all $ts \geq 2.76$, $p < .01$) higher mean ratings than those with MF values and, except for the color, sweater-type, and boot-height dimensions, MF values produced significantly (all $ts \geq 2.41$, $p < .05$) higher average ratings than LF values.

There were two small interactions, one for each group, in the analyses of variance of the test trial data. For the qualitative-dimension group, the value of sweater type had less effect on stimulus ratings when the LF value of fabric was present than when the HF or MF value of fabric was present [$F(4,76) = 3.20$, $p < .025$]. Similarly, when the LF value of boot height was present, the value of the number of buttons dimension had less effect [$F(4,76) = 2.65$, $p < .05$]. None of the other interactions was statistically significant.

The average ratings across subjects within a group indicate that all dimensions were used by some subjects in generating responses and that subjects were generally distinguishing between HF and MF values and between MF and LF values. For the group data, the rating difference between HF and MF values tended to be greater than the difference between MF and LF ratings, even in cases where both rating differences were statistically significant (the two exceptions were for the number of pockets and number of buttons dimensions). The prototype model predicts that dimension values equally distant from the dimension mean should produce equal ratings. As can be seen in Table 1, the HF and MF values are equally distant from the mean and should therefore

Table 3
Mean Test Trial Stimulus Ratings as a Function of Stimulus Dimension Values

Value	Dimension: Color	Fabric	Pants	Sweater	Jacket Length	Boot Height	Number of Buttons	Number of Pockets
HF	3.370	3.315	3.385	3.211	3.322	3.176	3.183	3.448
MF	2.809	2.909	2.863	2.880	2.857	2.846	2.924	3.007
LF	2.631	2.587	2.563	2.720	2.580	2.737	2.652	2.304
F(2,38)	24.23	26.02	22.41	8.69	29.37	7.94	18.26	68.77
SED	.111	.101	.124	.120	.098	.115	.088	.098

Table 4
Illustrative Data: Average Ratings and Rank Orderings of Dimension Values For Four Subjects

	Qualitative Dimension Group: Dimension			
	Color	Fabric	Pants	Sweater
Subject 1: MSe = .336				
HF Value	3.481	3.333	3.148	2.963
MF Value	2.519	2.704	2.815	2.778
LF Value	2.259	2.222	2.296	2.519
F(2,16)	33.28	24.92	14.79	4.00
p <	.001	.001	.001	.05
Ranking*	HF > MF = LF	HF > MF > LF	HF = MF > LF	HF = MF = LF
Subject 12: MSe = .012				
HF Value	3.000	3.333	3.333	3.333
MF Value	3.000	2.333	2.370	2.370
LF Value	2.037	2.370	2.333	2.333
F(2,16)	675.97	702.97	702.97	702.97
p <	.001	.001	.001	.001
Ranking*	HF = MF > LF	HF > MF = LF	HF > MF = LF	HF > MF = LF
	Numeric Dimension Group: Dimension			
	Jacket Length	Boot Height	Number of Buttons	Number of Pockets
Subject 5: MSe = .230				
HF Value	4.037	3.333	3.111	3.556
MF Value	3.259	3.185	3.148	3.593
LF Value	2.111	2.889	3.148	2.259
F(2,16)	110.23	6.01	.05	67.70
p <	.001	.025		.001
Ranking*	HF > MF > LF	HF = MF > LF		HF = MF > LF
Subject 8: MSe = .000				
HF Value	3.000	3.000	3.000	3.000
MF Value	2.000	2.000	2.000	2.000
LF Value	2.000	2.000	2.000	2.000
p <	.001	.001	.001	.001
Ranking*	HF > MF = LF	HF > MF = LF	HF > MF = LF	HF > MF = LF

**Based on t tests using MS_{Error} to estimate the standard error of the difference and with p = .05.*

produce equal ratings. The results of the present study are clearly inconsistent with this prediction when the data analysis is at the level of group means.

Individual Subject Analyses

When the data from individual subjects are examined, a somewhat more complicated picture appears. Not every subject attended to all four stimulus dimensions and the relative desirability of HF, MF, and LF values could vary across dimensions for any given subject. Four examples of data for individual subjects are given in Table 4, and summaries of the performance of individual subjects are given in Tables 5 and 6.

The data in Tables 4, 5, and 6 were compiled from analyses of variance on each subject's test trial data, with stimulus dimensions as the only factors in each analysis. The error term for all F tests was the four-way interaction, but essentially identical results were obtained by pooling all interactions for use as an error term. On the basis of these analyses, each subject could be typified by the number of stimulus dimensions for which the subject's responses varied significantly ($p < .05$) as a function of the dimension value present in the stimulus being rated. Most subjects attended to three of four stimulus dimensions, an average of 3.1 for the qualitative-dimension group and an average of 2.9 for the group with numeric dimensions. The distribution of subjects attending to four, three, two, and one dimensions for the qualitative-dimension group was, respectively, 11, 4, 2, and 2; one subject did not appear to be using any dimension as a basis of responding. The distribution for the numeric-dimension group was 6, 8, 4, and 2, with all subjects attending to at least one stimulus dimension. The data presented in Tables 5 and 6 did not appear to vary as a function of number of stimulus dimensions utilized by a subject, so only the overall data are presented.

Recall that the acquisition stimuli were selected so that there was pairwise independence for stimulus dimensions, that is, there was no information in the two-way interactions and the three-way interactions were minimal. For the 40 individual subject analyses, there were a total of 400 possible two- and three-way interactions. Only 29 of the interactions were significant ($p < .05$). The modal number of interactions per subject was 0, the median number per subject was 0, the mean

number per subject was .725, and the maximum number observed for any one subject was 5. It seems safe to conclude that, except for an occasional subject with some extra-experimental bias toward a particular combination of dimension values, the decision rules used by subjects involved an additive function of ratings for individual dimensions.

The four subjects selected for Table 4 are not necessarily typical of all 40 subjects. The data for these four subjects do, however, clearly indicate the complexity of the processing which subjects attempted and the dangers and difficulties present in making generalizations about concept formation and utilization processes. For example, Subject 1 of the qualitative-dimension group attended to all four stimulus dimensions in generating stimulus ratings but used the frequency differentials of the values differently across dimensions. It is difficult to say anything about the decision rule used by this subject other than that it was sensitive to frequency differentials.

In contrast, Subject 12 of the qualitative group presents an interesting case of a very orderly decision rule but one which is inconsistent with any simple decision models. If all 81 ratings given by Subject 12 are examined, one can see that, with the exception of one response (probably the result of a lapse of attention), the following procedure generates the ratings: Consider the HF values of fabric, pants, and sweater along with both the HF and MF values of color as "desirable" values and then calculate the rating for a stimulus by counting the number of desirable values present and adding one to the count. Subject 12 did not verbalize this procedure on the postexperimental questionnaire and may have been unable to verbalize it, but it describes his protocol exactly.

Subject 8 in the numeric-dimension group was very much like Subject 12 but with three important differences. First, the relative desirability of HF, MF, and LF values was consistent across dimensions. Second, Subject 8 made no "errors" in using her rule. Third, she was able to verbalize her rule. "First I set up a model of 33-9-4-3 as perfect or a rating of 5. If there was one deviance (sic) from this model rating (sic) was 4, two deviances (sic) 3, three–2, four–1. If all four were different from the model, this was what the public would dislike most." There was one subject in the qualitative-dimension group who behaved just like Subject 8, and there were several other subjects in both groups who were similar, both in the type of rule (but with less precision in its application) and in their ability to verbalize it.

The last subject included in Table 4, Subject 5 of the numeric-dimension group, is like Subject 1 of the qualitative group, except that he did not attend to all four dimensions. The number of buttons dimension was almost completely ignored, and there is an interesting shift in relative desirability from the jacket length dimension to the number of pockets dimension. Jacket length is a continuous variable, yet Subject 5 did not use a central tendency as a standard since the HF value produced higher stimulus ratings than the MF value. On the other hand, the number of pockets dimension is an integer variable and it would appear that 2.5 pockets is an ideal number for this subject. Subject 5 was only one of several subjects in the numeric-dimension group who had this type of protocol.

Strategies used by subjects. In spite of the degree of within- and between-subject variability in the way ratings were generated, some definite patterns appeared in the group data. First, when subjects were required to choose an ideal dimension value for each dimension following the test trials, 77.5% of the choices by subjects in the qualitative-dimension group (an average of 3.1 dimensions per subject) were of HF values. Subjects in the numeric-dimension group chose the HF value on an average of 2.25 dimensions (56.25%). Both of these means were significantly greater than the chance value of 1.33 (33.3%) [$t(19) = 8.68, p < .001$ and $t(19) = 3.40$, $p < .005$ for the qualitative and numeric groups, respectively]. This finding lends some degree of individual subject validation to the finding, reported above, that HF values received a mean rating on the test trials that was higher than the ratings for either MF or LF values. One additional point about these data: the average number of HF values chosen differed significantly for the two groups [$t(38) = 2.51$, $p < .025$]. This was the only measure on which the two groups differed significantly in the experiment and there is no ready explanation for this finding.

In examining the data for the individual subjects, two criteria were observed for giving consideration to an ordering of dimension values obtained from average test trial rating of uniform descriptions. First, as noted above, an ordering was included in the analysis only if there was adequate evidence that the subject was incorporating the dimension value in the response generation process, viz., only if the dimension produced statistically significant rating variance. For the qualitative-dimension group, 77.5% of the 80 orderings met this criterion, with 72.5% meeting the criterion for the numeric-dimension group. The second criterion for including an ordering in the analyses summarized in Tables 5 and 6 was that ordering of HF, MF, and LF values be reasonable given the acquisition trials and test trials with feedback. This criterion was that the HF value receive a higher rating than the LF value ($HF > LF$) and that the MF value receive an average rating which was neither significantly ($p < .05$ on a t test) higher than the HF value rating nor significantly lower than the rating of the LF value ($HF \geq MF \geq LF$). Of the orderings that met the first criterion, approximately 86% in the qualitative-dimension group and approximately 93% in the numeric group met the second criterion. The data of Tables 5 and 6 are based on the more than

Table 5
Distributions of Rank Orderings of Dimension Values For Dimensions With Significant Rating Variance and For Which the Values Were Ranked With HF > LF and HF ≥ MF ≥ LF

	Ordering	Source of Ordering: Raw Rating	Source of Ordering: Significant Rating
Group Q	HF = MF > LF	.148	.310
	HF > MF = LF	.227	.430
	HF > MF > LF	.625	.260
Group N	HF = MF > LF	.118	.456*
	HF > MF = LF	.140	.368*
	HF > MF > LF	.742	.140*

**For two subjects, the rating of the MF value did not differ significantly from the rating of either the HF or the LF value, despite significant rating variance on that dimension and a significantly higher rating for the HF as compared to the LF value. Thus, the entries in this column do not sum to 1.0.*

50 orderings for each group that met these criteria.

An additional problem arose in classifying each ordering. Subjects were not, for the most part, perfectly consistent in generating ratings. Thus, in many cases there was some doubt as to whether the subject was discriminating between two dimension values. Different types of errors in classifying rank orderings of dimension values are introduced by using nominally, but statistically nonsignificant, differences in mean rating and, alternatively, orderings based on only the rating differences that are statistically significant. Examples of the nature of this problem can be seen in Table 4. Because no good solution to this problem is apparent, the first column (Raw Rating) for each group in Tables 5 and 6 is based on orderings using the numeric differences (without regard to size of the differences) in ratings, while the second column (Significant Rating) is based on orders produced by statistically significant differences.

The data in Tables 5 and 6 permit three basic conclusions. First, the three different order types, HF > MF = LF, HF = MF > LF, HF > MF > LF, occur with approximately equal frequency. The data in Table 5 do not demand this conclusion, but, in the light of the problems in classifying orders noted above, the conclusion seems reasonable.

The second conclusion is based on the data in Table 6. The majority of subjects produced more than one ordering type. It would take a very strange pattern of errors in classifying orders to alter this conclusion.

Finally, the type of dimension, qualitative or numeric, has no effect on either the overall incidence of order types or upon the proportion of subjects producing more than one order type. The data in Table 5 support this conclusion [largest $t(38) = 1.367$, $p > .10$], as do the data in Table 6 ($\chi^2 < 1$ for both raw ratings and significant ratings).

Questionnaire data. The data from the questionnaires filled out by subjects paralleled the data presented in Table 5. The questionnaire allowed subjects to give two dimension values an equal rating but it is not clear that subjects always took advantage of this possibility. Of the 160 orderings produced by subjects, 75% of the orderings had HF > LF. Of these orderings, 86% had HF > MF, 79% had MF > LF, and 65% had HF > MF > LF. Note that the percentages are not directly comparable to the data in Table 5 since, in the case of the questionnaire data, the category HF > MF includes, in addition to the ordering HF > MF = LF, the two orderings HF > MF > LF and the ordering HF > LF > MF. Presumably, some of the examples of orderings of the latter two types reduce to HF > MF = LF if consideration is given to the possibility that a subject may not have used the procedure for representing ties. A similar caution is in order for the category MF > LF.

DISCUSSION

The data from the stimuli used for the test trials with feedback indicate that subjects were using a conceptual decision rule which was based primarily on information from the acquisition exemplars. It is unlikely that subjects learned very much in a rote fashion from the first two blocks of nine test trials with feedback. With the exception of the 4HF and 0HF stimuli, each of the 16 stimuli used for the test trial with feedback blocks appeared only once during the first two blocks. Ratings of these 16 stimuli during the final block of test trials with feedback and during the final test series exhibited a stable pattern and level of performance: very clear discrimination between HF and LF values on all four stimulus dimensions and low but much greater than chance accuracy in making the "correct" response. A conceptual mode of responding based upon the relative frequency of HF and LF values in the acquisition exemplars is more compatible with this data than a rote recall mode.

The results of the present study are incompatible with the view that more than a very few subjects, if any, used a decision rule based on distance of the to-be-rated stimulus from either a prototype or the exemplars presented during acquisition (the average distance model). The basic problem for all distance models is that type of dimensions, qualitative or numeric, had no effect on the types of decision rules used by subjects. In order for distance models to predict this absence of a dimension effect, it is necessary to make two assumptions: that subjects can make distance judgments on the dimensions used in the qualitative group and that the psychological distances between values on the qualitative dimensions are ex-

Table 6
Proportions of Subjects Producing One, Two, or Three Types of Rank Orderings

	Source of Ordering	Number of Types: 1	2	3
Group Q	Raw Rating	.50	.44	.06
	Significant Rating	.33	.61	.06
Group N	Raw Rating	.53	.42	.05
	Significant Rating	.37	.53	.10

tremely similar to the corresponding psychological distances on the numeric dimensions. While it is clear that subjects can judge the similarity to two values on a qualitative dimension, a procedure equivalent to making a distance judgment, it is also clear that subjects can make these judgments using a variety of psychological dimensions. For example, it is possible to compare the similarity of mohair, wool, and nylon on the basis of durability, softness, cost, frequency of usage in clothing, washability, etc. The similarity of any two values on this fabric dimension depends completely on the particular psychological dimension the subject selects for a basis of judgment. Thus, to predict an equivalence between qualitative- and numeric-dimension results, it is necessary to assume that the correct proportion of subjects chose an appropriate combination of psychological dimensions which had the correct similarity relationships between values. We feel that such a large number of generally dubious assumptions is inappropriate when a viable alternative exists. We prefer to simply assume that subjects were using only frequency of values in generating ratings.

It could be argued that models based on distance judgments should be applied only to situations where the stimulus dimensions have a compelling quantitative characteristic which the great majority of subjects would use in making judgments. The stimuli for the numeric group seem to be such a case. However, prototype models which use the mean as a measure of central tendency encounter a serious problem with the data of this group. Such a prototype model predicts that HF and MF values should produce equivalent ratings. The data clearly contradict this prediction; an average of 50% to 88% (cf., Table 5) of the dimension value orderings produced by subjects in the numeric dimension group rated HF values higher than MF values.

In all of the above calculations and discussion, it has been assumed that subjects would use the objective values of the numeric dimensions in calculating a mean. We are aware that some researchers believe that scaled dimension values and some other measure of central tendency should be utilized in assessing the descriptive power of distance models. We agree that decades of psychological research and many recent studies have shown the need for psychological scales and the power of multidimensional scaling techniques in recovering psychological spaces. Indeed, it is likely that the "fit" of a distance model to the group data and to some individual subject data could be improved by these procedures. On the other hand, it is not clear what would be gained by the application of these procedures to the data of this experiment. Individual subjects clearly behave differently than the group data would lead us to believe, some subjects clearly are not using a decision function based on distance from a measure of central tendency (e.g., Subject 8 in Table 4), and using scaled dimension values still does not address the issue of the equivalence of the data from numeric and qualitative groups. It seems to us that it is much more parsimonious and accurate to simply point out that subjects are generally sensitive to differences in the frequencies of occurrence of values and they adopt a variety of decision functions utilizing frequency differentials and probably, in some cases, dimension values.

Individual subjects were unwilling or unable to utilize all of the information they possessed. From the questionnaires, it was quite apparent that most subjects knew that the three stimulus values on each stimulus dimension occurred with different frequencies. However, to map reliably the 15 stimulus-equivalence classes generated by all combinations of HF, MF, and LF values on the four dimensions into five response classes is an arduous, if not impossible, task. Not one of the 40 subjects did it. Subjects solved the problem by reducing the number of stimulus-equivalence classes in one or the other of two ways, by ignoring some stimulus dimensions or by ignoring frequency differentials within dimensions. Notice that the rule used by Subject 8 in Table 4—count the number of HF values and add one to generate the correct response—works perfectly for the test trials with feedback of Table 2. As Table 6 indicates, the majority of subjects were not consistent across dimensions in how they reduced the stimulus class-to-response mapping problem. The variables responsible for this shift in strategy are not known. Stimulus dimension type, however, was not one of them in the present study, nor was particular dimension within a type.

The results of Reed (1972) and Reed and Friedman (1973) appear to be in conflict with those of the current study. However, there are a number of possible reasons for the differing results. First, there are methodological differences, including those cited earlier which involve characteristics of stimulus dimension distributions. Furthermore, Reed used a classification task rather than a rating task, as was used in the present study. While rating of a stimulus with respect to both categories is not required as an operation that is logically prior to classification, the models which fit the Reed (1972) and Reed and Friedman (1973) data assume the subject rates a stimulus with respect to each of the available categories as a basis for classification. Thus, it does not seem the task difference is the most promising basis for reconciliation.

Another difference between the Reed (1972) and Reed and Friedman (1973) studies and the present one is in the data analysis. All of Reed's conclusions are based on analyses of group data. The group data for the present study, as displayed in Table 3, are completely in accord with a frequency model, which assumes ratings are a simple function of stimulus dimension value frequencies. Such a model predicts that HF values should produce higher ratings than MF values and that MF values should, in turn, produce higher ratings than LF values. In addition, these group data are clearly incompatible with a prototype-distance model, predicting that HF and MF stimulus values should produce equal ratings. But, as has been stressed above, not one of the 40 subjects responded in complete agreement with the

frequency model and several, for example, Subjects 8 and 12 in Table 4, were in direct conflict with it. The group data do not represent the individual subject data; they are, in fact, a potentially serious distortion of individual subject data. Given the unusual dimension-value distributions, the extra-experimental biases, the varying objective dimensional validities, and other characteristics of the Reed (1972) and Reed and Friedman (1973) studies, it is possible that their subjects used a mixture of strategies which fortuitously produced averaged group data compatible with distance models.

At the outset, we noted an apparent lack of progress in extending our knowledge beyond Attneave's (1957) summary. In one sense, the results of the present study serve to accentuate our ignorance. Still, there are some optimistic notes. First, it is possible to examine in detail an individual subject's performance in probabilistic concept learning tasks and make some sense out of it. The individual subject is more complicated than we typically allow in our theories, but there is order. Second, the results of the present study demand that, in studying the acquisition and utilization of concepts, we attend not only to what is learned but to how the learned information is used. Subjects in the present study clearly knew much more than they utilized in rating stimuli. Finally, Attneave's (1957) evaluation of what is learned specifies the notion of a central tendency prototype or schema. The results reported here indicate that subjects do not necessarily learn a central tendency. It is indeed possible for the subject to calculate one if it is required or advantageous, from a record of frequencies (probabilities or strength) per stimulus value. In fact, there is no evidence that subjects have anything in memory before they begin rating stimuli except the memories of some of the exemplars they have experienced. It is entirely possible that subjects do not normally judge test stimuli against a prototype or schema, frequency or central tendency, but against the raw memory events, as with the average distance model (Reed, 1973). While it is beyond the scope of this paper to prove this assertion, we claim that there is no evidence of the necessity of a schema to account for the data of any concept learning experiment currently in the literature. A frequency model which assumes every act of remembering is a conceptual inference based on independent episodic memory traces will handle the data adequately.

REFERENCE NOTE

1. Hayes-Roth, F., & Hayes-Roth, B. *A schematic model of abstraction.* (Tech. Rep. MMPP 74-2). Ann Arbor, Michigan: University of Michigan, Department of Psychology, December 1973.

REFERENCES

Aiken, L. S., & Griffin, L. R. Visual and auditory processing of common pattern class structure. *Perception & Psychophysics*, 1972, **12**, 492-496.

Attneave, F. Transfer of experience with a class-schema to identification-learning of patterns and shapes. *Journal of Experimental Psychology*, 1957, **54**, 81-88.

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, **1**, 360-368.

Beach, L. R. Cue probabilism and inference behavior. *Psychological Monographs*, 1964, **78**, 5(Whole No. 582). (a)

Beach, L. R. Recognition, assimilation, and identification of objects. *Psychological Monographs*, 1964, **78**, 6(Whole No. 583). (b)

Bourne, L. E., Jr. An inference model for conceptual rule learning. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium*. Hillsdale, N.J: Erlbaum, 1974.

Bransford, J. D., & Franks, J. J. The abstraction of linguistic ideas. *Cognitive Psychology*, 1971, **2**, 331-350.

Dansereau, D. F., & Brown, B. R. The attribute selection process in pattern perception: The effect of constraint redundancy and stimulus exposure time on the classification of spatially represented Markov patterns. *Memory & Cognition*, 1974, **2**, 75-81.

Franks, J. J., & Bransford, J. D. Abstraction of visual patterns. *Journal of Experimental Psychology*, 1971, **90**, 65-74.

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

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

Keele, S. W. *Attention and human performance*. Pacific Palisades, Calif: Goodyear, 1973.

Levine, M. *A cognitive theory of learning: Research on hypothesis testing*. Hillsdale, N.J: Erlbaum, 1975.

Millward, R. B., & Wickens, T. D. Concept-identification models. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 1). San Francisco: W. H. Freeman, 1974.

Neumann, P. G. An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, 1974, **2**, 241-248.

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

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

Posner, M. I. *Cognition: An introduction*. Glenview, Ill: Scott, Foresman, 1973.

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

Reed, S. K. *Psychological processes in pattern recognition*. New York: Academic Press, 1973.

Reed, S. K., & Friedman, M. P. Perceptual vs. conceptual categorization. *Memory & Cognition*, 1973, **1**, 157-163.

Reitman, J. S., & Bower, G. H. Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, 1973, **4**, 194-206.

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

(Received for publication August 16, 1977; revision accepted January 7, 1978.)