Journal of Experimental Psychology:
Learning, Memory, and Cognition
1984, Vol. 10, No. 1, 104–114

Choice, Similarity, and the Context Theory of Classification

Robert M. Nosofsky
Harvard University

Medin and Schaffer's (1978) context theory of classification learning is interpreted in terms of Luce's (1963) choice theory and in terms of theoretical results obtained in multidimensional scaling theory. En route to this interpretation, quantitative relationships that may exist between identification and classification performance are investigated. It is suggested that the same basic choice processes may operate in the two paradigms but that the similarity parameters that determine performance change systematically according to the structure of the choice paradigm. In particular, when subjects are able to attend selectively to the component dimensions that compose the stimuli, the similarity parameters may tend toward what is optimal for maximizing performance.

Medin and Schaffer (1978) proposed a quantitative model, termed the *context theory*, to account for subjects' performance in classification paradigms. The context theory has provided good fits to data in numerous classification experiments (Medin, 1982; Medin, Altom, Edelson, & Freko, 1982; Medin, Dewey, & Murphy, 1983; Medin & Schaffer, 1978; Medin & Smith, 1981). The theory assumes, essentially, that subjects' classification of a given stimulus is determined on the basis of its similarity to the stored category exemplars. The purpose of this article is to provide an interpretation of the context theory in terms of previous work concerning choice and similarity. This interpretation places the context model within a well-established theoretical framework and allows one to ask new questions and extend the model in ways that were not previously obvious. In addition, it may help in attempts to gain a fuller understanding of quantitative relationships that may exist between subjects' identification and classification performance.

The work reported in this article was supported by Grants BNS 80-26656 from the National Science Foundation and MH 37208 from the National Institute of Mental Health to Harvard University.

I would like to thank J. Busemeyer, R. D. Luce, and D. L. Medin for helpful comments and suggestions regarding earlier versions of this article. I especially wish to thank W. K. Estes and E. E. Smith for numerous useful suggestions and discussions, and for their encouragement.

Requests for reprints should be sent to Robert M. Nosofsky, Department of Psychology and Social Relations, Harvard University, Cambridge, Massachusetts 02138.

In the empirical work for which the context model has been verified thus far, the stimuli have varied along four binary-valued dimensions, and the subjects have been required to classify the stimuli as belonging in one of two categories. To simplify the discussion, this basic structure is assumed in the initial part of this article; in addition, the following notation is employed:

1. Let X and Y denote the two categories.
2. Let $\mathbf{x}$ denote a stimulus in category X, and let $\mathbf{y}$ denote a stimulus in category Y.
3. Let x_j denote the value of stimulus $\mathbf{x}$ on dimension j, and similarly for y_j.
4. Let $\mathbf{t}$ denote a given test stimulus, and let t_j denote the value of stimulus $\mathbf{t}$ on dimension j.
5. Let $S(\mathbf{t}, \mathbf{x})$ denote the similarity of stimulus $\mathbf{t}$ to stimulus $\mathbf{x}$.

As is clear from Medin and Schaffer's (1978, pp. 211–214) discussion, the context theory states that the probability of classifying test stimulus $\mathbf{t}$ as a member of category X, $P(X|\mathbf{t})$, is given by

$$P(X|\mathbf{t}) = \frac{\sum_{\mathbf{x}\in X} S(\mathbf{t}, \mathbf{x})}{\sum_{\mathbf{x}\in X} S(\mathbf{t}, \mathbf{x}) + \sum_{\mathbf{y}\in Y} S(\mathbf{t}, \mathbf{y})}. \quad (1)$$

The individual $S(\mathbf{t}, \mathbf{x})$s are computed by the following multiplicative rule:

$$S(\mathbf{t}, \mathbf{x}) = \prod_{j=1}^{4} s_j, \quad (2)$$

where $s_j = p_j$, $(0 \leq p_j \leq 1)$, if $t_j \neq x_j$; and

$s_j = 1$, if $t_j = x_j$. The individual p_js are the parameters to be estimated in the model.

Response Rule

As is rather evident, the response-ratio rule (Equation 1) proposed by Medin and Schaffer (1978) bears a striking structural resemblance to Luce's (1963) choice model for stimulus identification. According to that theory, the probability of stimulus i leading to response j in an identification experiment, $P(R_j|S_i)$, is given by

$$P(R_j|S_i) = \frac{\beta_j \eta_{ij}}{\sum_k \beta_k \eta_{ik}}, \qquad (3)$$

where $0 \leq \beta_j, \eta_{ij} \leq 1$, $\Sigma\, \beta_j = 1$, $\eta_{ij} = \eta_{ji}$, and $\eta_{ii} = 1$. The β_j parameters are interpreted as response-bias parameters, and the η_{ij} parameters are interpreted as similarity measures on the stimuli S_i and S_j. The index k in the denominator of Equation 3 ranges over the set of stimuli that are eligible as responses in the experiment.

Medin and Schaffer's (1978) response rule may be viewed as a bias-free extension of Luce's (1963) choice theory applied at the category level, by simply defining $\eta(\mathbf{t}, X) = \Sigma_{\mathbf{x}\in X}\, \eta(\mathbf{t}, \mathbf{x})$. A natural question that arises, however, is whether or not there is some theoretical account of the relationship $\eta(\mathbf{t}, X) = \Sigma_{\mathbf{x}\in X}\, \eta(\mathbf{t}, \mathbf{x})$.

A simple account of this relationship may reside in the corresponding qualitative structures of the identification and classification paradigms. The one-to-one mapping of stimuli onto responses in identification is transformed into a many-to-one mapping of stimuli onto responses in classification. An obvious starting hypothesis for a quantitative model relating the two paradigms, originally proposed by Shepard, Hovland, and Jenkins (1961), and Shepard and Chang (1963), may be stated as follows: To predict classification performance from identification performance, one should simply cumulate over all stimulus–response cells in the identification matrix that would map onto a given stimulus–response cell in the classification matrix. Thus, any interstimulus confusion in the identification matrix which is a within-category confusion would result in a correct classification response. I will refer to this as the *mapping hypothesis.*

If the mapping hypothesis were correct, then the response rule proposed by Medin and Schaffer (1978) could be derived directly from Luce's (1963) choice theory applied at the stimulus identification level. The probability of making response X given stimulus i in the classification paradigm, $P(R_X|S_i)$, is found by summing over the probabilities that a stimulus in category X is chosen in the identification paradigm:

$$\begin{aligned} P(R_X|S_i) &= \sum_{j\in X} P(R_j|S_i), \\ &= \sum_{j\in X} \left(\frac{\beta_j \eta_{ij}}{\sum_k \beta_k \eta_{ik}}\right), \\ &= \left(\sum_{j\in X} \beta_j \eta_{ij}\right)\Big/\left(\sum_k \beta_k \eta_{ik}\right). \qquad (4) \end{aligned}$$

In a bias-free experiment, this is the response rule proposed by Medin and Schaffer (1978).

The main problem with this theoretical account is that the mapping hypothesis was explored systematically by Shepard et al. (1961) and rejected on empirical grounds. These researchers demonstrated convincingly that one could not predict subjects' classification performance from their identification performance on the same set of stimuli. We are left, therefore, with the following paradox: Luce's (1963) choice theory and Medin and Schaffer's (1978) context theory provide good fits to data in identification and classification experiments, respectively. The mapping hypothesis provides an obvious theoretical link between the two theories, but it has been rejected empirically. I attempt to resolve this paradox later in this article. The argument to be advanced is that the similarity parameters that determine performance may change systematically as one goes from identification to classification instructions. First, however, a theoretical interpretation of Medin and Schaffer's multiplicative similarity rule (Equation 2) is provided.

Multiplicative Similarity Rule[1]

The second aspect of the context theory to be considered is the multiplicative rule for

[1] It was brought to my attention by D. L. Medin (personal communication, September 14, 1982) that Takane and Carroll (1982) have pointed out the same functional re-

computing stimulus similarity (Equation 2). This has been deemed by Smith and Medin (1981) as the crucial feature of the context theory that differentiates it from previous theories of classification learning. The main question to be addressed is whether the multiplicative rule for computing stimulus similarity can be meaningfully integrated with previous findings in the psychological literature regarding stimulus similarity. Although an intuitive rationale for the multiplicative rule is provided by Medin and Schaffer (1978) and Medin and Smith (1981), little mention is made of any independent empirical support for the rule in general, or of its relationship to previously proposed models regarding stimulus similarity.

One of the major assumptions made by researchers in the field of multidimensional scaling theory is that stimulus similarity is some monotonically decreasing function of psychological distance. Letting D denote psychological distance between two stimuli and letting f be some monotonically decreasing function, then using the previous notation we have

$$S(\mathbf{t}, \mathbf{x}) = f[D(\mathbf{t}, \mathbf{x})], \tag{5}$$

that is, the similarity between test stimulus **t** and stimulus **x** is some monotonically decreasing function of the psychological distance between stimulus **t** and stimulus **x**. For various measurement–theoretic reasons (Beals, Krantz, & Tversky, 1968), the distance metric between two stimuli is generally assumed to take the form of the Minkowski-r metric:

$$D(\mathbf{t}, \mathbf{x}) = \left(\sum_{j=1}^{n} |t_j - x_j|^r\right)^{1/r}, \tag{6}$$

where $r \geq 1$, n is the number of dimensions composing the given stimuli, and t_j and x_j are appropriately scaled psychological values. The particular value of r which yields a best fit to the similarity data has been found to depend heavily on the type of dimensions composing a stimulus. In particular, the city-block metric ($r = 1$) is appropriate for stimuli having separable dimensions, and the Euclidean metric ($r = 2$), for those with integral dimensions.[2] Intuitively, separable dimensions are highly analyzable and remain psychologically distinct when in combination. In contrast, integral dimensions combine into relatively unanalyzable, integral wholes. Various other converging operations are used to distinguish between these two types of dimensions (Garner, 1974).

The stimulus sets used by Medin and Schaffer (1978) in their initial experiments were composed of stimuli that varied in form, color, size, and number. On the basis of previous work these would clearly be considered separable dimensions. Thus, the psychological distance between test stimulus **t** and stimulus **x** is given by

$$D(\mathbf{t}, \mathbf{x}) = \sum_{j=1}^{n} d_j, \tag{7}$$

where $d_j = |t_j - x_j|$, appropriately scaled. Note that because the dimensions were binary-valued, each d_j is equal either to 0 (if $t_j = x_j$), or to q_j, $0 \leq q_j < \infty$ (if $t_j \neq x_j$).

Reviewing our previous equations, we have

$$S(\mathbf{t}, \mathbf{x}) = \prod_{j=1}^{n} s_j, \quad 0 \leq s_j \leq 1,$$

according to the multiplicative rule of Medin and Schaffer, and we have

$$S(\mathbf{t}, \mathbf{x}) = f\left(\sum_{j=1}^{n} d_j\right), \quad 0 \leq d_j < \infty,$$

according to the basic assumptions of multidimensional scaling theory. The question I now raise is whether there is some function f which produces the multiplicative rule posited by

lationships that I note in this section. The work that I report here was carried out independently of these researchers' work.

[2] The city-block and Euclidean metrics might best be considered as approximations to the best-fitting Minkowski-r metric for stimuli composed of separable and integral dimensions. Tversky and Gati (1982) argued that a large amount of similarity data for stimuli composed of separable dimensions is best fit by values of r somewhat less than 1. (When r is less than 1, of course, the Minkowski-r "metric" is no longer a distance metric at all, since the triangle inequality is not satisfied; this does not invalidate a spatial approach to modeling stimulus similarity, only the metric assumption.) For mathematical convenience I assume that $r = 1$ serves as a fair approximation to the best-fitting Minkowski-r for stimuli composed of separable dimenions. Of course, an improved theory may result by using values of r less than 1.

Medin and Schaffer (1978). The answer is that the multiplicative rule arises if and only if

$$f(x) = e^{-cx}, \quad c > 0. \tag{8}$$

That is, we have

$$e^{-c\sum_{j=1}^{n} d_j} = \prod_{j=1}^{n} e^{-cd_j}, \tag{9}$$

and because $0 \leq e^{-cd_j} \leq 1$, we simply set $s_j = e^{-cd_j}$. The case of $|t_j - x_j| = 0$ (identical values) maps onto $s_j = 1$, and $|t_j - x_j| = \infty$ onto $s_j = 0$. A proof that this functional relationship is unique may be found in Roberts (1979).

In summary, the basic hypothesis being proposed is that the multiplicative rule of computing stimulus similarity arises as a special case of psychological distance between stimuli conforming to the city-block metric, and of stimulus similarity being an exponential decay function of psychological distance.

The second part of this hypothesis receives some independent support from earlier work by Shepard (1957, 1958a, 1958b) regarding the relationship between stimulus generalization (measured in terms of confusion errors in absolute identification tasks) and psychological distance. He concluded that the relation was well described by an exponential decay function on the basis of both empirical data and more primitive underlying theoretical assumptions. Luce (1963) also incorporated the assumption that similarity is an exponential decay function of psychological distance in the development of his choice theory.

More recently, Getty, Swets, Swets, and Green (1979) utilized this assumption with a great deal of success in predicting subjects' confusion errors in an identification task from their similarity ratings of the same stimuli. First, they applied a multidimensional scaling procedure to the similarity judgments to construct a psychological space and to obtain the locations of the stimuli in that space. In predicting the subjects' confusion errors on the subsequent identification task, they then employed Luce's (1963) choice model with the assumption that stimulus similarity is an exponential decay function of psychological distance. The fit of the model to data was excellent.

Optimization of Similarity Relations

In summary, Medin and Schaffer's (1978) context theory arises as a logical consequence of integrating the mapping hypothesis of the identification–classification relationship with the following well-established models in the areas of choice and similarity: (a) Luce's (1963) choice model for stimulus identification, (b) an exponential decay function relating stimulus generalization to psychological distance, and (c) psychological distance relationships for stimuli composed of separable dimensions conforming (approximately) to the city-block metric. Furthermore, since the context model provides good fits to data in numerous experiments, the framework seems to hold together very well. The paradox in this formulation, however, is that the mapping hypothesis has been explored systematically by Shepard et al. (1961) and rejected on empirical grounds. To resolve this discrepancy it is worthwhile to examine the results of Shepard et al. in some detail.

Shepard et al. (1961) studied the relationship between identification and classification performance for sets of eight stimuli that varied along three binary-valued separable dimensions. In the classification tasks, each set was divided into two categories of four stimuli each. In general, given any eight unique stimuli, there are 70 distinct ways to partition them into two groups of four. However, the Shepard et al. (1961) investigation was simplified by the fact that, for stimuli varying along three binary-valued dimensions, the 70 partitions fall into six distinct types. The partitions within each type are structurally equivalent in the following sense: Any 2 partitions that are of a given type can be transformed into one another by a reassignment of dimension roles. (For example, a partition that is formed by placing all black stimuli in category X and all white stimuli in category Y is structurally equivalent to a partition that is formed by placing all large stimuli in category X and all small stimuli in category Y. Note that both of these partitions are structured according to a single stimulus dimension; in the first case the relevant dimension is color, whereas in the second case the relevant dimension is size.) The six basic classification types are illustrated schematically in Figure 1. So, for example, for

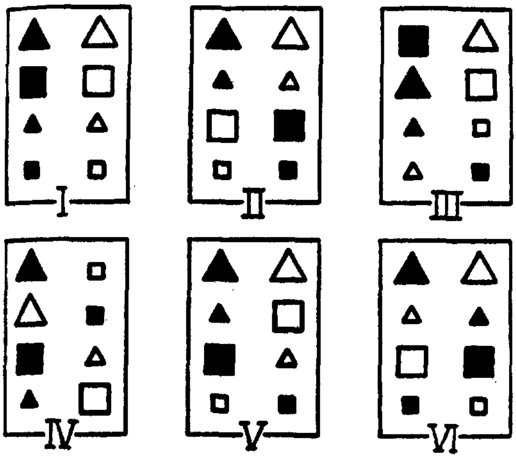

Figure 1. Schematic illustration of the six classification types investigated by Shepard et al. (1961). (Within each box the four stimuli on the left belong in one class and the four stimuli on the right in the other class. From "Learning and memorization of classifications" by R. N. Shepard, C. I. Hovland, and H. M. Jenkins, 1961, *Psychological Monographs, 75,* 13, Whole No. 517, p. 3. Copyright 1961 by the American Psychological Association. Reprinted by permission.)

the Type II partition (see Figure 1), the subject should respond X if the stimulus is triangular and black, or if it is square and white; otherwise, the subject should respond Y. Another example of a Type II partition, with interchanged dimension roles, would be the following: "Respond X if the stimulus is triangular and large, or if it is square and small," and so forth.

A graph-theoretic representation of the six classification types that provides further insight into their structure is presented in Shepard et al. (1961, p. 4). For present purposes, the reader should note that only one dimension is relevant for performing a Type I classification, and two dimensions are relevant for performing a Type II classification. The Type VI classification is an extreme case in which all three dimensions are equally relevant. Types III, IV, and V are intermediate in structural complexity between Types I and II and Type VI, with all three dimensions being relevant, but to varying extents. These three types may be described as *single dimension plus exception* classifications, with the precise nature of the exception varying among the types. For example, for the Type V classification, all black stimuli are assigned to category X, and all white stimuli to category Y, except that the small, white square is switched with the small, black square.

Shepard et al. (1961) ran a series of experiments in which subjects were required to learn each of the classification types illustrated in Figure 1. Prior to these classification tasks they ran an identification condition in which subjects were required to learn a unique response to each stimulus. In both the identification condition and the various classification conditions the procedure used was the standard paired-associate learning paradigm, in which a stimulus was presented, the subject made a response, and the correct response was then provided by the experimenter. Learning on each condition continued for 400 trials, or until a subject reached a criterion of 32 consecutive correct responses. The total number of errors made during the learning of each classification type was recorded. Shepard et al. then used what is essentially the mapping hypothesis to predict subjects' classification performance from their identification performance, and compared the predicted total number of errors for each classification type with the observed total number of errors.[3] The results are presented in Figure 2. As is evident, the simple mapping hypothesis fails. First, the predicted number of errors exceeded the observed number of errors, with the magnitude of the discrepancies varying among the different types. Second, the observed amount of variation in number of errors among the types exceeded the predicted amount of variation. And third, the observed rank order of difficulty between Type II and Types III–V was reversed from what was predicted.

There are a couple of plausible explanations for the discrepancies in the predicted and ob-

[3] A precise specification of the method of prediction would go beyond the scope of the present article. In its essentials, however, it was an application of the mapping hypothesis; namely, one predicts classification errors from identification errors by cumulating over all stimulus confusions in the identification condition that would result in between-category confusions in the classification condition. The reader should note that although subjects participated in six classification conditions, the predictions are all made from a single identification condition. One cumulates over different cells in the identification matrix for each particular classification prediction.

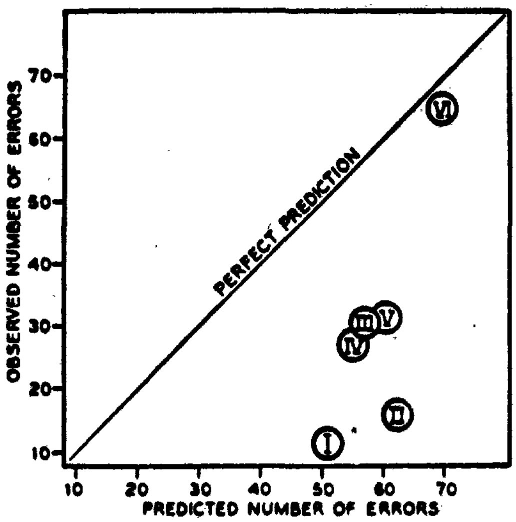

Figure 2. Observed number of errors made during the learning of each type of classification, plotted against the number of errors predicted from the identification learning results. (From "Learning and memorization of classifications", by R. N. Shepard, C. I. Hovland, and H. M. Jenkins, 1961, *Psychological Monographs, 75*, 13, Whole No. 517, p. 27. Copyright 1961 by the American Psychological Association. Reprinted by permission.)

served patterns of performance. A partial answer might be that there was a generalized increase in sensitivity in the classification tasks relative to the identification one. This might have occurred for various reasons, including the fact that all the classification tasks were conducted subsequent to the identification one. A more compelling explanation, however, suggested by Shepard et al. (1961), is that some process of selective attention was involved. To perform a Type I classification, for instance, a subject need attend only to one dimension—color in the example in Figure 1. In an identification task, on the other hand, all three dimensions are relevant and the subject must divide attention. Note that in a Type VI classification, all the dimensions are equally relevant, and that the classification results are fairly well predicted from the identification results in this case.

Medin and Schaffer (1978) noted that selective attention might play a central role in determining the similarity parameter values in the context model. Therefore, if the focusing of attention is influenced by the structure of the categories, as suggested earlier, then it follows that the similarity parameters that determine performance may depend systematically on the structure of the categories as well. Indeed, Medin (1982) has begun to investigate aspects of category structure that might be relevant in this regard.

These lines of reasoning may be modeled within the current theoretical framework in a straightforward way. We have assumed that the similarity between stimuli **t** and **x** is given by

$$S(\mathbf{t}, \mathbf{x}) = e^{-(d_1+d_2+\cdots+d_n)}, \qquad (10)$$

where $d_j = q_j$, $(0 \le q_j < \infty)$, if $t_j \ne x_j$; and $d_j = 0$, if $t_j = x_j$. An alternative way of writing this equation is

$$S(\mathbf{t}, \mathbf{x}) = e^{-D(w'_1+w'_2+\cdots w'_n)}, \qquad (11)$$

where $D = \Sigma_{j=1}^{n} q_j$, $w'_j = q_j/D = w_j$, if $t_j \ne x_j$; and $w'_j = 0$, if $t_j = x_j$. Note that $0 \le w_j \le 1$ and $\Sigma\, w_j = 1$. The D parameter may be interpreted as total distance in psychological space, and is analogous to a sensitivity parameter, whereas the w_j parameters are weights assigned to each dimension in computing overall distance. Assuming that the component dimensions are of roughly equal perceptual salience, then the w_j parameters model the effect of attending selectively to the component dimensions.[4]

Now, further assuming that the context-model response rule (Equation 1) does indeed govern subjects' classification probabilities, then it is possible to determine for any given value of D the distribution of dimension weights that optimizes performance in a given classification paradigm. (By *optimize*, here, I mean maximize the average percentage correct.) And it is reasonable to hypothesize that, with learning, subjects will distribute attention among the component dimensions in a way that tends to optimize performance. This approach was used by Getty et al. (1979) and Getty, Swets, and Swets (1980) in predicting

[4] R. M. Shiffrin (personal communication, March 16, 1983) suggested that the D parameter may be interpreted as consisting of two components, one associated with general sensitivity and the other with some form of resource sharing. According to this view, there may be some relationship between the value of D and the distribution of weights. This is an interesting possibility which awaits future investigation.

stimulus–response confusion matrices in identification experiments. They obtained some support for the hypothesis that, with learning, subjects' weights of component dimensions tended toward optimum. I now use the same approach in an attempt to model the identification–classification relationship observed by Shepard et al. (1961).

The context model's predictions of the average percentage of errors for each classification type investigated by Shepard et al. (1961) are plotted in Figure 3. These predictions assume two different combinations of D (overall sensitivity) and **w** (distribution of attention). I use the percentage of errors here, rather than the percentage correct, to facilitate comparison with Shepard et al.'s results in Figure 2. In case the method of prediction is not evident to the reader, a brief example is provided in the Appendix. The purpose of plotting these theoretical predictions against one another is to gain an abstract characterization of the identification–classification relationship observed by Shepard et al. To the extent that the relationships among the theoretical predictions mirror the empirical relationships in Figure 2, support for the present modeling account is obtained. The reader should note, however, that the present formulation is intended to serve only as an approximation to the qualitative pattern of results in Figure 2. The model used here is a static one, whereas Shepard et al.'s results emerged from a dynamic learning process. A complete quantitative account of those results would undoubtedly need to specify the way in which D and **w** change with experience.

The predictions plotted against the horizontal axis in Figure 3 assume a uniform distribution of weights, that is, $w_1 = w_2 = w_3$, with $D = 3$. Assuming that subjects distribute attention equally among the component dimensions in the identification task, then these predictions are analogous to predicting classification performance from identification performance by directly applying the mapping

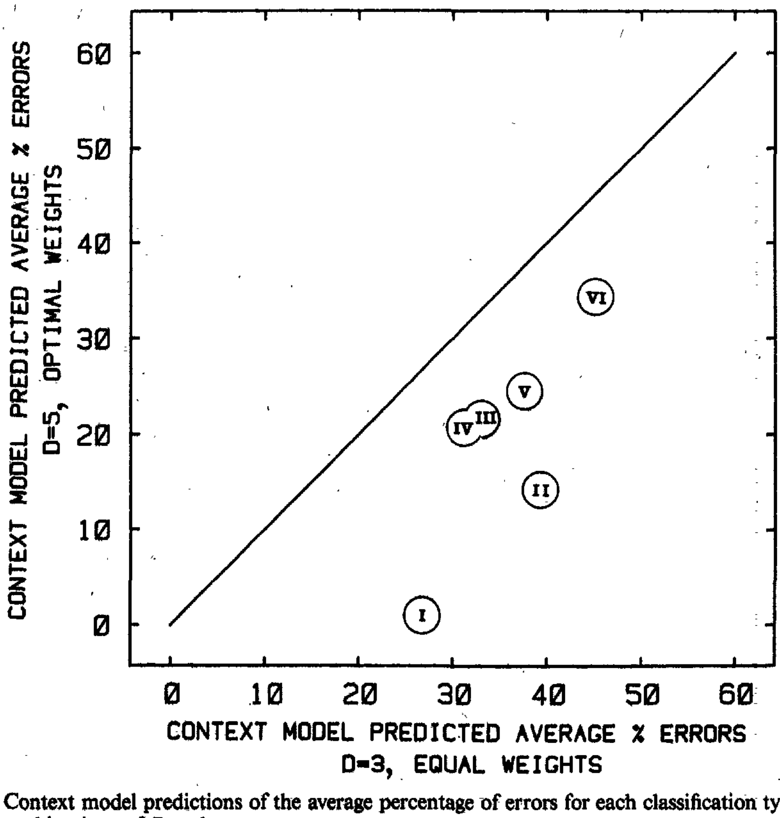

Figure 3. Context model predictions of the average percentage of errors for each classification type for two different combinations of D and **w**.

hypothesis, except that the predictions are now being made by a mathematical model, rather than by actual cumulation over observed identification results. The predictions plotted against the vertical axis in Figure 3 assume an optimal distribution of weights for each classification type and assume that overall sensitivity has increased from $D = 3$ to $D = 5$. The increase in the value of D represents the generalized increase in sensitivity that might have occurred in the classification conditions because they were conducted subsequent to the identification condition; this assumption is not crucial, however, for bringing out the relationships seen in Figure 3. The particular values of D chosen in Figure 3 are arbitrary, and similar patterns of results emerge for a wide range of D values. The distributions of weights that are optimal for each classification type are summarized in Table 1.[5]

Table 1
Distributions of Weights That Are Optimal for Each Classification Type

Classification type	Optimal dimension weights 1	2	3
I	1.00	0.00	0.00
II	0.50	0.50	0.00
III	0.35	0.35	0.30
IV	0.33	0.33	0.33
V	0.46	0.27	0.27
VI	0.33	0.33	0.33

Note. Dimension weights that optimize the average percentage correct for each classification type, subject to the constraint that $D = 5$. Dimensions 1, 2, and 3 correspond to color, shape, and size, respectively, as illustrated in Figure 1. The optimal dimension weights for Types III and V will vary depending on the value of D; the optimal dimension weights for Types I, II, IV, and VI are stable.

The representation in Figure 3 appears to reflect fairly well the qualitative pattern of results in Figure 2. First, the pattern of improvement for each classification type, relative to what is predicted by a direct application of the mapping hypothesis, agrees fairly systematically with the findings of Shepard et al. (1961): A good deal of improvement occurs for Type I and Type II classifications, less for Types III–V, and still less for Type VI.[6] Second, the amount of variation in number of errors among the types is increased with optimal weights as compared with equal weights. Third, and perhaps most important, the predicted order of difficulty for Type II and Types III–V also parallels Shepard et al.'s results: Performance for Types III–V exceeds that for Type II with equal weights, but the order reverses with optimal weights. This reversal was a highly emphasized result in Shepard et al.'s theoretical discussion. The reader may note in addition that if Shepard et al.'s findings on the III-IV-V triad are taken to be perfectly reliable, then the present formulation makes a perfect prediction of the rank order of difficulty for the six types, both the order predicted from the identification results (equal weights) and the observed classification results (optimal weights).

In summary, the interpretation of the context model that is offered here is capable of mirroring the qualitative pattern of results that Shepard et al. (1961) observed when they compared identification and classification performance. As stated previously, the context-model response rule arises as a logical consequence of integrating the mapping hypothesis with Luce's (1963) choice model for stimulus identification. According to the formulation developed here, however, the mapping relation that links identification and classification performance may not be a direct

[5] The weights (w) that are optimal for each classification type may be determined by the classical methods of constrained optimization; namely, one solves for the w_j and λ that maximize the Lagrangian function

$$L(\mathbf{w}, \lambda) = \sum_{\mathbf{x}\in X} P(X|\mathbf{x}, D, \mathbf{w}) + \sum_{\mathbf{y}\in Y} P(Y|\mathbf{y}, D, \mathbf{w}) + \lambda(1 - \sum w_j),$$

(where $0 \le w_j \le 1$) by using standard methods of the calculus (see Daellenbach & George, 1978). For some classification types, however, the computation is difficult. Therefore, one uses a computerized parameter-search routine to find the w_js that maximize the average percentage correct for each classification type.

[6] Perhaps the major shortcoming of the representation in Figure 3 is that performance on the Type VI classification is predicted to be substantially better than what is predicted by a direct application of the mapping hypothesis (whereas Shepard et al., 1961, observed little improvement). Actually, if there were no increase in general sensitivity (D), then there would be no predicted improvement for the Type VI classification, because the optimal distribution of weights for this type is the uniform one. Unfortunately, there would also be little predicted improvement for the III-IV-V triad in this circumstance.

one, but a more abstract linkage. In particular, the similarity parameters that determine performance may change systematically according to the structure of the choice paradigm that is investigated. A good working hypothesis is that the similarity parameters tend toward what is optimal for maximizing performance.

The optimal pattern of dimension weights being posited here is an attempt to quantify the processes of selective attention and abstraction that may mediate identification and classification performance. Medin et al. (1983) have demonstrated, however, that the process of abstraction does not occur automatically, but depends crucially on experimental conditions. One experiment with particularly favorable conditions was reported by Medin et al. (1983, last-name-infinite condition). Subjects were required to classify photographs of women's faces into two categories. The experimenters coded the faces as varying along four binary-valued dimensions: color of hair (light or dark), color of shirt (light or dark), type of smile (open or closed), and length of hair (long or short). They informed the subjects that these were the relevant dimensions for performing the classification. (The faces varied freely on all irrelevant dimensions.) The important manipulation in this experiment was that subjects were never presented with the same face twice; rather, a different face was used each time to instantiate the logical coding corresponding to each abstract stimulus.

It was expected that this manipulation would enhance the processes of selective attention and abstraction in subjects' classification learning for two reasons. First, because subjects would never be retested on any particular exemplar, the use of a rote paired-associate learning strategy would be discouraged. Under these experimental conditions, the use of a rote learning strategy would benefit subsequent performance far less than a strategy in which subjects attempted to abstract the relevant category structure. A second reason, closely related to the first, is that the potential size of the exemplar population that defined the categories in this experiment was essentially infinite. Various researchers have found that the process of abstraction improves with increases in category size (e.g., Homa, Cross, Cornell, Goldman, & Schwartz, 1973).

The category structure used by Medin et al. (1983) is presented in Table 2. The estimated similarity parameters that achieved a best fit of the context model to the classification data were $p_1 = .23$, $p_2 = .72$, $p_3 = .25$, and $p_4 = .24$. These similarity parameters may be transformed to distances (q_j), according to the present formulation, by setting

$$q_j = -\ln p_j.$$

Total distance in psychological space (D) is then given by

$$D = \sum_{j=1}^{4} q_j$$

and the weights corresponding to each dimension by

$$w_j = q_j/D.$$

The weights that are obtained by transforming the similarity parameters in this manner are plotted in Figure 4 (solid curve). Also plotted in Figure 4 are the weights that are theoretically optimal according to the present formulation (dashed curve). That is, these are the weights that would have maximized subjects' average percent-correct scores in this experiment. The close correspondence between the best-fitting weights and the theoretically optimal weights suggests that the current framework is worthy of further investigation.

Table 2
Category Structure Used by Medin et al. (1983)

Category exemplar	Dimension 1	2	3	4
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1
B1	1	1	0	0
B2	0	1	1	0
B3	0	0	0	1
B4	0	0	0	0

Summary

To summarize, in this article I attempted to relate Medin and Schaffer's (1978) context theory to a more general theoretical framework for the modeling of choice and similarity. The choice rule for classification was related to Luce's (1963) choice model for identification

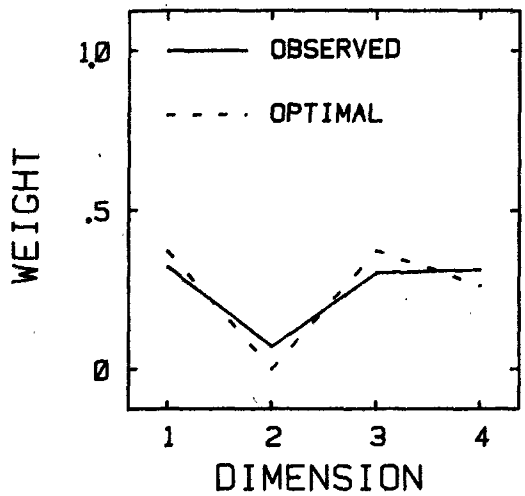

Figure 4. Observed and optimal dimension weights for the last-name-infinite condition of Medin et al. (1983) according to the present model. (The observed weights are computed from Medin et al.'s estimated similarity parameters, p_j, by setting $q_j = -\ln p_j$, and then setting $w_j = q_j/\Sigma q_j$. The optimal weights are computed according to the model as explained in Footnote 5, subject to the constraint that $D = \Sigma q_j$.)

by means of a hypothesized mapping relation between identification and classification performance. The multiplicative similarity rule was related to theoretical results obtained in multidimensional scaling research. Finally, it was suggested that the relation between identification and classification performance may be understood in terms of selective attention to the component dimensions that compose the stimuli. The hypothesis that subjects may distribute attention among the component dimensions in a way that tends to optimize performance was investigated, and gained some preliminary support. It is hoped that the present work may contribute to the development of a unified quantitative framework for understanding identification and classification performance.

References

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review, 75,* 127–142.

Daellenbach, H. G., & George, J. A. (1978). *Introduction to operations research techniques.* Boston: Allyn & Bacon.

Garner, W. R., (1974). *The processing of information and structure.* New York: Wiley.

Getty, D. J., Swets, J. B., & Swets, J. A. (1980). Use of perceptual dimensions in signal identification. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 361–380). Hillsdale, NJ: Erlbaum.

Getty, D. J., Swets, J. A., Swets, J. B., & Green, D. M. (1979). On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics, 26,* 1–19.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Schwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101,* 116–122.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Medin, D. L. (1982). Structural principles of categorization. In B. Shepp & T. Tighe (Eds.), *Interaction: Perception, development and cognition* (pp. 203–230). Hillsdale, NJ: Erlbaum.

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

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

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

Medin, D. L., & Smith, E. E. (1981). Strategies in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 241–253.

Roberts, F. S. (1979). *Measurement theory with applications to decisionmaking, utility, and the social sciences.* Reading, MA: Addision-Wesley.

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

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

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

Shepard, R. N., & Chang, J. J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology, 65,* 94–102.

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

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Takane, Y., & Carroll, J. D. (1982, June). Multidimensional scaling from the tree construction method of similarity data. Handout for a talk given at the joint meeting of the Psychometric Society and the Classification Society, University of Quebec, Montreal, Canada.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89,* 123–154.

(Appendix follows on next page)

Appendix

Context Model Predictions of Classification Performance

To predict the average percentage of errors on classification Type II assuming that $D = 3$ and $w_1 = w_2 = w_3$, that is, $\mathbf{w} = (1/3, 1/3, 1/3)$, one proceeds as follows. First, the average percentage correct (APC) on each classification type is given by

$$\mathrm{APC} = 100[\sum_{\mathbf{x} \in X} P(X|\mathbf{x}, D, \mathbf{w}) + \sum_{\mathbf{y} \in Y} P(Y|\mathbf{y}, D, \mathbf{w})]/8, \tag{A1}$$

where $P(X|\mathbf{x}, D, \mathbf{w})$ is given by Equations 1 and 11 in the text. Using the partition illustrated in Figure 1, Type II, as an example, and letting color, shape, and size correspond to Dimensions 1, 2, and 3, respectively, then the probability of classifying the large, black triangle (call it $\mathbf{x1}$) in its correct category is given by

$$P(X|\mathbf{x1}) = \frac{e^{-3(0+0+0)} + e^{-3(0+0+1/3)} + e^{-3(1/3+1/3+0)} + e^{-3(1/3+1/3+1/3)}}{[e^{-3(0+0+0)} + e^{-3(0+0+1/3)} + e^{-3(1/3+1/3+0)} + e^{-3(1/3+1/3+1/3)}] + [e^{-3(0+1/3+1/3)} + e^{-3(0+1/3+0)} + e^{-3(1/3+0+1/3)} + e^{-3(1/3+0+0)}]}, \tag{A2}$$

(i.e., the large, black triangle shares all its features with itself; shares color and shape with the small, black triangle; shares size with the large, white square; and shares no features with the small, white square; and so forth). The Type II classification has a very simple structure, and the reader may verify that $P(X|\mathbf{x})$ is equal to .607 for all $\mathbf{x} \in X$, and likewise for $P(Y|\mathbf{y})$. Thus, substituting into Equation A1 yields

$$\mathrm{APC} = 100[4(.607) + 4(.607)]/8 = 60.7. \tag{A3}$$

The average percentage of errors (APE) is then given by APE = 100 − 60.7 = 39.3.

Received July 29, 1982
Revision received June 14, 1983 ■