Memory & Cognition
1994, 22 (3), 352-369

Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961)

ROBERT M. NOSOFSKY
Indiana University, Bloomington, Indiana

MARK A. GLUCK
Rutgers University, New Brunswick, New Jersey

THOMAS J. PALMERI and STEPHEN C. McKINLEY
Indiana University, Bloomington, Indiana

and

PAUL GLAUTHIER
Rutgers University, New Brunswick, New Jersey

Memory & Cognition
1994, 22 (3), 352-369

Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961)

ROBERT M. NOSOFSKY
Indiana University, Bloomington, Indiana

MARK A. GLUCK
Rutgers University, New Brunswick, New Jersey

THOMAS J. PALMERI and STEPHEN C. McKINLEY
Indiana University, Bloomington, Indiana

and

PAUL GLAUTHIER
Rutgers University, New Brunswick, New Jersey

We partially replicate and extend Shepard, Hovland, and Jenkins's (1961) classic study of task difficulty for learning six fundamental types of rule-based categorization problems. Our main results mirrored those of Shepard et al., with the ordering of task difficulty being the same as in the original study. A much richer data set was collected, however, which enabled the generation of block-by-block learning curves suitable for quantitative fitting. Four current computational models of classification learning were fitted to the learning data: ALCOVE (Kruschke, 1992), the rational model (Anderson, 1991), the configural-cue model (Gluck & Bower, 1988b), and an extended version of the configural-cue model with dimensionalized, adaptive learning rate mechanisms. Although all of the models captured important qualitative aspects of the learning data, ALCOVE provided the best overall quantitative fit. The results suggest the need to incorporate some form of selective attention to dimensions in category-learning models based on stimulus generalization and cue conditioning.

Recent years have seen an avalanche of newly proposed models of category learning and representation. As such models grow increasingly more sophisticated, there is a need to develop increasingly more rigorous testing grounds so that one may choose among them. Most previous attempts to test alternative models have focused on the end products of categorization by observing patterns of transfer data following an initial learning phase. In the spirit of developing more rigorous tests, there has been a renewed interest in understanding details of the category *learning* process (see, e.g., Estes, 1986; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Nosofsky, Kruschke, & McKinley, 1992). Beyond simply predicting transfer data following the completion of category learning, the following question arises: How well can alternative models predict patterns of classification during the entire learning sequence?

The purpose of our study was to collect a rich set of classification learning data that would provide a useful testing ground for the numerous models that have been proposed. A seemingly infinite variety of learning paradigms are available, but we hoped to collect some learning data that researchers might regard as fundamental. Although the ultimate goal of categorization researchers is the development of a model that can account for all forms of classification phenomena, it seems worthwhile to focus initial efforts on primary and basic forms of classification learning data.

A classic study of category learning is the one reported by Shepard, Hovland, and Jenkins (1961), who studied the difficulty of learning six fundamental types of categorization problems. Their results proved to be highly diagnostic for ruling out various models of classification learning based solely on elementary principles of stimulus generalization and cue conditioning. As will be seen, their data continue to challenge current models of classification learning.

In the first part of our article, we will review the design and results of Shepard et al.'s (1961) elegant and in-

This work was supported by Grant PHS R01 MH48494-01 from the National Institute of Mental Health to R.M.N., an ONR Young Investigator Award to M.A.G., and Grant N00014-88-J-0112 from the ONR Biological Intelligence Initiative to M.A.G. The authors thank John Kruschke and three anonymous reviewers for their comments on an earlier version of this article. Correspondence concerning this article may be addressed to either R. M. Nosofsky, Department of Psychology, Indiana University, Bloomington, IN 47405 (e-mail: nosofsky@ucs.indiana.edu), or M. A. Gluck, Center for Molecular and Behavioral Neuroscience, Rutgers University, 197 University Ave., Newark, NJ 07102 (e-mail: gluck@pavlov.rutgers.edu).

fluential study. We will then report a partial replication and extension of that study. By using their basic paradigm, while collecting a more extensive data set, we should provide a fundamental testing ground for formal models of classification learning. Finally, we begin the testing process by quantitatively fitting three formal models to the observed learning data: Anderson's (1991) rational model, Kruschke's (1992) ALCOVE model, and an extended version of Gluck and Bower's (1988b) configural-cue model.

The Shepard et al. (1961) tasks are examples of "rule-based" category learning problems. Simple deterministic logical rules allow one to classify all exemplars with perfect accuracy. By contrast, the computational models that we test in this article were developed primarily to account for the learning of fuzzy, ill-defined category structures. Nevertheless, we argue that it is important to understand how these models fare with data from simpler rule-based classification tasks. To the extent that these models are adequate for learning both rule-based and ill-defined structures, we may presume that a single unified process accounts for both types of learning. To the extent that the models apply to only one class of tasks, we might infer the existence of multiple learning strategies. This would then give rise to further questions regarding how and why alternative learning strategies take precedence. Indeed, it is probably for the reasons just stated that current researchers have repeatedly reached back to the rule-based tasks of Shepard et al. (1961) as a canonical data base for evaluating models of human classification learning (Anderson, 1991; Estes, 1994; Gluck & Bower, 1988b; Kruschke, 1992; Nosofsky, 1984).

A Review of Shepard et al.'s (1961) Study

In Shepard et al.'s (1961) study, subjects were tested on six basic types of classification problems. In each problem, there were eight stimuli constructed from three binary-valued dimensions. Four of the stimuli belonged to one category, and the other four stimuli to a second category. These constraints result in six problem types, which are illustrated by the cubes in Figure 1. The vertices of the cubes represent individual stimuli. The oval vertices represent stimuli assigned to Category 1, and the rectangular vertices represent stimuli assigned to Category 2. Each face of a cube represents a value along one of the binary-valued dimensions. For ease of description, we imagine that the dimensions correspond to shape

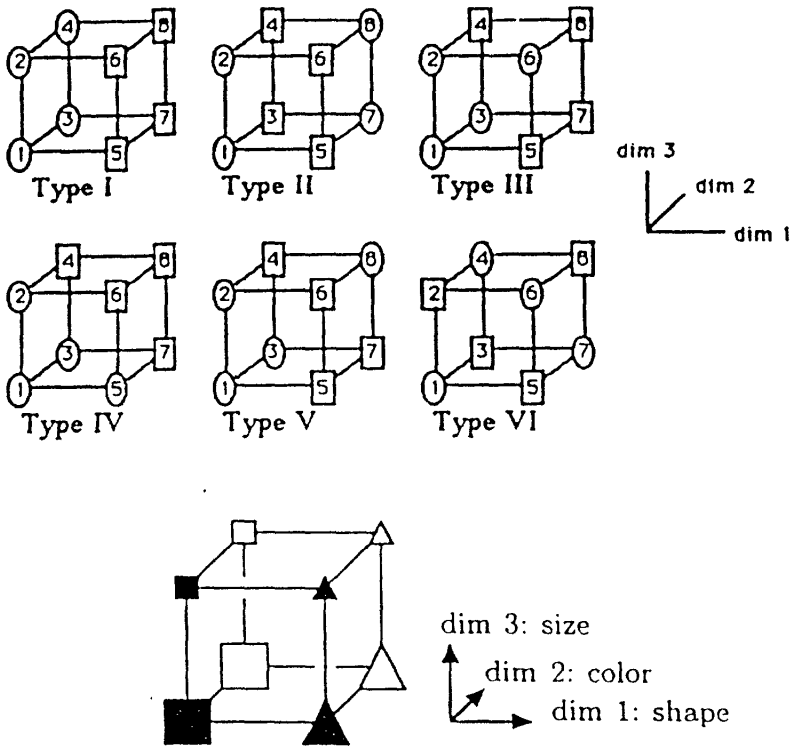

Figure 1. Top: The six types of categorization problems tested by Shepard, Hovland, and Jenkins (1961). The eight stimuli are denoted by the corners of the cubes. Assignments to categories are denoted by the ovals or rectangles that enclose the stimulus numbers. Bottom: Illustrative example in which the stimuli vary along the dimensions of shape (square vs. triangle), color (black vs. white), and size (large vs. small).

(square vs. triangle), color (black vs. white), and size (large vs. small), as is illustrated in the bottom part of Figure 1. Any assignment of stimuli to categories, with four stimuli in each of the two categories, can be rotated or reflected onto one of the six structures shown in the figure.

The simplest category structure is the Type I problem. Here, information about only one dimension (shape in the example in Figure 1) is necessary to solve the problem. For Type II, exactly two dimensions are relevant. In the Figure 1 example, black squares and white triangles are assigned to Category 1, whereas white squares and black triangles are assigned to Category 2. Information about size is irrelevant to solving the problem. Note that Type II is the exclusive-or problem along its two relevant dimensions. The Type VI problem is the most complex category structure, with all three dimensions being equally relevant to solving the problem. Stating a logical rule for Type VI in terms of values on each of the component dimensions amounts to enumerating the stimuli in each of the categories.[1] Finally, Types III, IV, and V are intermediate in structural complexity between Type II and Type VI. All three dimensions are relevant, but to differing extents. These three types can be described as "single dimension plus exception" structures. For example, for Type V, squares are assigned to Category 1 and triangles to Category 2, except the small white square is switched with the small white triangle. We discuss the subtle structural differences among the Type III, IV, and V problems later in our article.

Shepard et al. (1961) measured the difficulty of learning these problem types in terms of the number of errors that subjects made until they reached a criterion. They found that Type I was the easiest classification to learn, followed by Type II, followed by Types III, IV, and V, which were about equal in difficulty, and finally Type VI. (This ordering of difficulty pertains to subjects' initial encounters with each problem type, not to some more intricate transfer effects that were also observed in their studies.)

Of particular interest was that the Type II problem was learned with fewer errors than were Types III, IV, and V, whereas a variety of models based on stimulus generalization and cue conditioning predicted the opposite ordering of difficulty. Recall that in the Type II problem only two dimensions are relevant, whereas in Types III–V all three dimensions are relevant. Shepard et al. (1961) suggested that for the Type II problem, subjects may have learned to focus attention on the two relevant dimensions, whereas for Types III–V, subjects had to spread attention across all three dimensions. Such a process might account for the observed ordering of difficulty. Support for this interpretation was provided in their original study and in additional theoretical analyses of Shepard et al.'s data reported by Nosofsky (1984) and Kruschke (1992). This theme of the role of selective attention in classification learning serves to highlight some of the model comparisons that we report later in our article.

Goals of Our Replication and Extension

Despite the elegance of its design and its profound influence, Shepard et al.'s (1961) seminal investigation had some limitations. First, data were collected from only 6 subjects. To test the reliability and generalizability of the results, it is important to conduct a similar study with a larger sample. Second, Shepard et al. reported their data in terms of total number of errors observed for each problem type. A richer and more intricate data set is yielded by observing the errors made for each problem type at different points during the learning sequence. In other words, it is important to track the actual time course of learning instead of recording only the cumulative errors made by the end of learning. Third, the error data were reported for each problem type as a whole. For several of the problems, individual items within each category have distinct statuses. Some of the items within these problem types are expected to be easier to learn than others. Tracking the time course of learning for individual items within each problem type might provide still more useful information. The goal of our research, therefore, was to replicate Shepard et al.'s paradigm, but collect a richer data set by (1) testing more subjects, (2) tracking the time course of learning for each problem, and (3) studying the difficulty of learning individual items within various of the problem types.

Finally, although previous researchers have discussed the ability of different models to account for qualitative aspects of Shepard et al.'s (1961) data, in this research we begin the process of quantitatively testing such models. By meeting our goals of testing more subjects, generating block-by-block learning curves, and studying the difficulty of learning individual item types, we produce a rich set of classification learning data that is suitable for quantitative fitting and that allows for rigorous comparisons among the alternative models.

METHOD

Subjects

The subjects were 120 undergraduates from Indiana University, who participated as part of an introductory psychology course requirement.

Stimuli and Apparatus

The stimuli were geometric forms with lines that filled their interiors. The stimuli varied along three binary-valued dimensions: shape (squares or triangles), type of interior lines (solid or dotted), and size (large or small). These stimuli are fairly representative of the types of separable-dimension stimuli used in Shepard et al.'s (1961) original studies. The stimuli were presented on the screens of CompuAdd 320 computers, and the subjects entered their responses on the computer keyboards.

Procedure

The logical structure of the six problems that were tested is shown in the top part of Figure 1. Assignment of physical dimensions and of the values on each dimension to this logical structure was randomized for each subject and problem that was tested.

Each of the 120 subjects was tested on two classification problems, for a total of 40 subjects per problem. All pairs of problems

were tested equally often, and the order of problems within each pair was balanced across subjects. The subjects were given explicit instructions that the relevant rule and dimensions for the second problem were chosen independently of those that were relevant in the first problem.

The procedure for the learning of each problem was similar to the one used by Shepard et al. (1961). In the first and second block of 8 trials, each stimulus appeared once in a random order. In each subsequent block of 16 trials, each stimulus appeared twice in a random order. (This procedure directly follows the one used by Shepard et al.) On each trial, a stimulus appeared on the screen, the subject classified it into Category 1 or 2, and feedback was provided. Learning continued until a subject reached a criterion of 4 consecutive sub-blocks of 8 trials with no errors, or for a maximum of 400 trials (25 blocks of 16 trials).

RESULTS

The average probabilities of errors for each problem in each block of 16 trials are reported in Table 1 and are shown graphically in Figure 2. Note that the means on late blocks reflect zero values for subjects who had already reached criterion. Our assumption is that the subjects who had reached criterion, and who thereby had already achieved between 32 and 40 consecutive correct responses, would have continued to respond without error if they maintained the same level of motivation and concentration.

Table 1
Average Error Proportions for Each of Problem Types I–VI in Blocks 1–25 of Learning

	Problem Type					
Block	I	II	III	IV	V	VI
1	0.211	0.378	0.459	0.422	0.472	0.498
2	0.025	0.156	0.286	0.295	0.331	0.341
3	0.003	0.083	0.223	0.222	0.230	0.284
4	0.000	0.056	0.145	0.172	0.139	0.245
5	0.000	0.031	0.081	0.148	0.106	0.217
6	0.000	0.027	0.078	0.109	0.081	0.192
7	0.000	0.028	0.063	0.089	0.067	0.192
8	0.000	0.016	0.033	0.063	0.078	0.177
9	0.000	0.016	0.023	0.025	0.048	0.172
10	0.000	0.008	0.016	0.031	0.045	0.128
11	0.000	0.000	0.019	0.019	0.050	0.139
12	0.000	0.002	0.009	0.025	0.036	0.117
13	0.000	0.005	0.008	0.005	0.031	0.103
14	0.000	0.003	0.013	0.000	0.027	0.098
15	0.000	0.002	0.009	0.000	0.016	0.106
16	0.000	0.000	0.013	0.000	0.014	0.106
17	0.000	0.000	0.008	0.000	0.014	0.078
18	0.000	0.000	0.006	0.000	0.014	0.077
19	0.000	0.000	0.009	0.000	0.013	0.078
20	0.000	0.000	0.003	0.000	0.014	0.061
21	0.000	0.000	0.005	0.000	0.013	0.058
22	0.000	0.000	0.000	0.000	0.009	0.042
23	0.000	0.000	0.003	0.000	0.011	0.042
24	0.000	0.000	0.005	0.000	0.008	0.030
25	0.000	0.000	0.002	0.000	0.008	0.038
Average	0.010	0.032	0.061	0.065	0.075	0.143

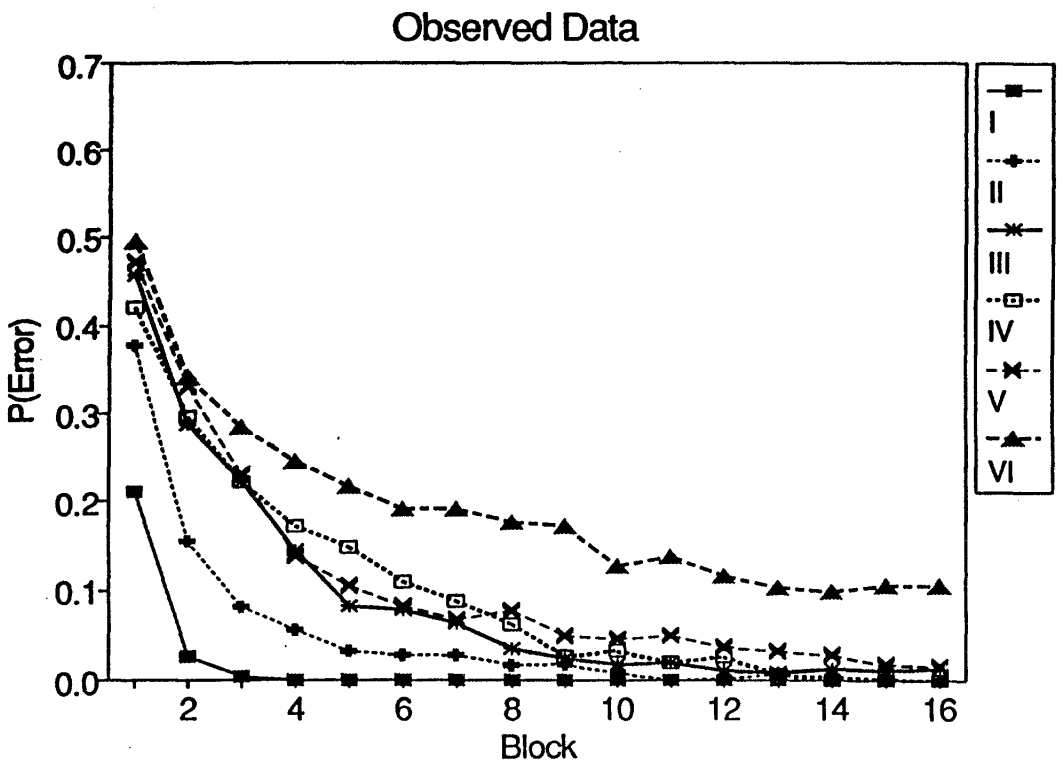

Figure 2. Average probabilities of errors for each problem in each block of 16 trials. The data from only Blocks 1–16 are shown. There were essentially zero errors during Blocks 17–25 for all problems except Type VI (see Table 1).

As can be seen, the main results closely replicate those of Shepard et al. (1961). The fewest errors occurred for the Type I problem, followed by Type II, followed by Types III, IV, and V, which were about equal in difficulty, and finally Type VI. Pairwise t tests showed that Type I was learned with significantly fewer errors than was Type II [$t(78) = 4.353$, $p < .01$]; that Type II was learned with significantly fewer errors than were Types III, IV, and V [average $t(78) = 3.515$, $p < .01$]; and that Types III, IV, and V were learned with significantly fewer errors than was Type VI [average $t(78) = 3.227$, $p < .01$]. There were no significant differences among Types III, IV, and V. Figure 3 shows the data separately for the first and second problems learned. As can be seen, there was an overall practice effect, with fewer average errors on the second problem than on the first problem. However, the overall ordering of difficulty for Types I-VI was the same for the first and second problems learned, and statistical tests yielded the same results as for the pooled data.[2]

The average number of trials to criterion for Problem Types I-VI was 44.0, 85.4, 121.6, 127.0, 133.8, and 189.2, respectively. These results mirror the ones for average number of errors, with Type I learned more quickly than Type II; Type II learned more quickly than Types III, IV, and V; and Types III, IV, and V learned more quickly than Type VI. Across all 240 subject-problem combinations (120 subjects × 2 problems each), there were only 6 cases of a subject failing to reach the learning criterion: 4 in Type VI, 1 in Type III, and 1 in Type V.

In Problem Types I, II, and VI, all individual items have the same logical status, in the sense that their roles within the category structure are all logically the same. However, in Types III, IV, and V, all items do not have the same structural roles, and some may be easier to learn than others.

In Type IV, for example, Stimuli 1 and 8 can be characterized as central members of their categories, whereas the remaining stimuli are peripheral members—see Figure 1. (Recall that the Type IV structure can be described by a single-dimension-plus-exception rule. In the Type IV problem, a central member is one that always participates in the single-dimension rule and is never considered an exception, whereas peripheral members will sometimes serve as exceptions, depending on which dimension is used for the rule.) Likewise, in the Type III problem, Stimuli 1, 2, 7, and 8 can be described as central members and Stimuli 3, 4, 5, and 6 as peripheral members. And in the Type V problem, there are three distinct types of items: Stimuli 1 and 5 are central, Stimuli 2, 3, 6, and 7 are peripheral, and Stimuli 1 and 8 are exceptions. (The exceptions are the stimuli that violate the only single-dimension rule that is available for the Type V problem.)

We recorded detailed learning curves for each of these item types in Problems III, IV, and V and found that, for all problems, the central members were learned with fewer errors than the peripheral members were, whereas the exceptions had the most errors of all. Because the formal models that we test subsequently in our article all successfully predict this qualitative pattern of results, and the quantitative fits of the models to these data turned out not to be diagnostic, we will not consider the individual item-type data further.[3]

OVERVIEW OF THREE FORMAL MODELS OF CLASSIFICATION LEARNING

In this section, we test three quantitative models on their ability to predict our classification learning data: Kruschke's (1992) ALCOVE model, Anderson's (1991) rational model, and an elaborated version of Gluck and Bower's (1988b) configural-cue model. These three models stand among the leading models of classification learning in the field today, with each model being able to characterize a wide variety of fundamental classification phenomena (for reviews, see Anderson, 1990, 1991; Gluck, Bower, & Hee, 1989; and Nosofsky & Kruschke, 1992). To date, however, there have been few attempts to develop quantitative tests to compare these models. This goal of comparing the ability of the models to quantitatively predict fundamental sets of learning data was the primary motivation for the present study.

Because ALCOVE and the rational model have been discussed in detail in several previous articles (e.g., Anderson, 1990, 1991; Kruschke, 1992; Nosofsky & Kruschke, 1992), we will summarize them here only briefly. We will provide a more extended discussion of the elaborated configural-cue model.

A common assumption made when fitting all three models is that the stimuli are composed of three binary-valued dimensions, as illustrated by the structures in Figure 1. In other words, we assume that the psychological dimensions that compose the objects correspond directly to the physical dimensions. Furthermore, because assignment of physical dimensions to the logical structures that define each category was randomized in our experiment, the intrinsic salience of each logical dimension is assumed to be equal. Although some logical dimensions are more diagnostic than others for determining category membership, it is the models' job to *learn* these diagnosticities.

ALCOVE

ALCOVE is an extended version of the well-known context model of classification (Medin & Schaffer, 1978; Nosofsky, 1986). It extends the context model by placing it in a connectionist framework and providing it with the learning mechanisms found in adaptive networks (Gluck & Bower, 1988a; Rumelhart, Hinton, & Williams, 1986). According to ALCOVE, people represent categories by storing individual exemplars in memory, and they form associations between these exemplars and the categories to be learned. Exemplars are represented as points in a multidimensional psychological space. When an object is presented, it activates each exemplar according to its similarity to that exemplar, with similarity a de-

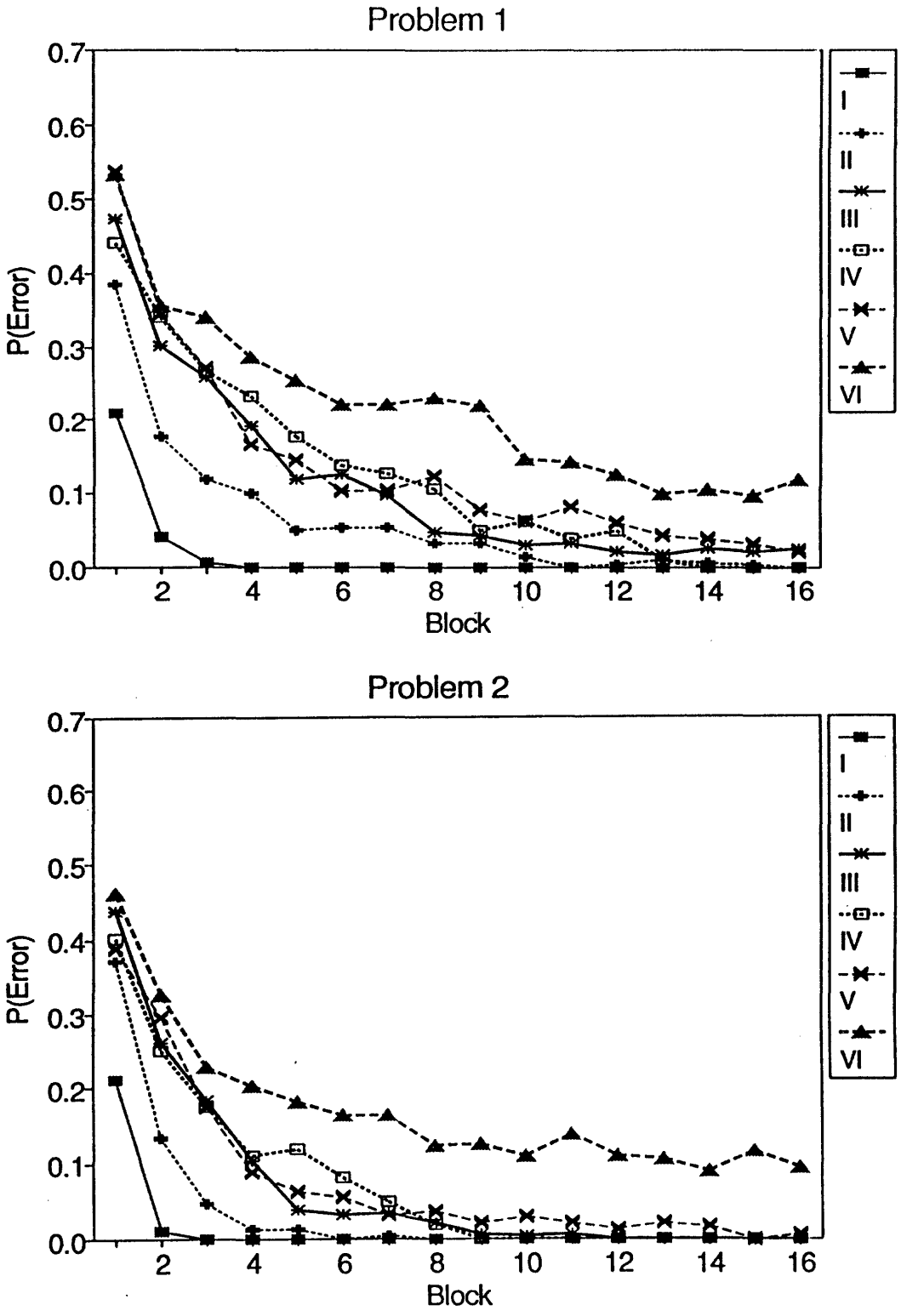

Figure 3. Average probabilities of errors for each problem in each block of 16 trials, shown separately for the first and second problems learned.

creasing function of distance in the space (Shepard, 1987). The extent to which each category is activated is determined jointly by the similarity of the object to the stored exemplars and by the strength of association between the exemplars and the alternative categories. For example, if an object is highly similar to a set of exemplars that are strongly associated to category A, then, when that object is presented, category A will be strongly activated.

An important component assumption of ALCOVE is that selective attention processes can modify similarities among objects in the multidimensional space (Nosofsky, 1986). Selective attention is represented by a set of weights that "stretch" and "shrink" the space along its component dimensions. In general, ALCOVE learns to attend selectively to the dimensions that are relevant to solving a classification problem and to ignore the dimensions that are irrelevant.

Formally, ALCOVE is a three-layered, feed-forward connectionist network as illustrated in Figure 4. The input nodes code the values on the psychological dimensions that compose the stimuli. The hidden nodes represent locations in the multidimensional space in which the exemplars are embedded. Each hidden node corresponds to a unique training exemplar. The output nodes code the degree to which each category is activated. The input nodes are gated to the hidden exemplar nodes by dimensional attention weights, and the hidden exemplar nodes are connected to the output nodes by a set of association weights.

When object i is presented, each exemplar node j is activated by the function

$$a_{ij} = \exp[-x\sum\alpha_m|x_{im}-x_{jm}|], \quad (1)$$

where x_{im} and x_{jm} are the values of exemplars i and j on dimension m (constrained to be either 0 or 1 for the present binary-valued stimuli illustrated in Figure 1); x is an overall sensitivity parameter that scales distances in the psychological space; and α_m is the (learned) attention weight on dimension m. This activation function assumes a city-block metric for computing distance in psychological space and an exponential decay function for transforming distance into similarity (Shepard, 1987). Thus, the more similar an exemplar is to an input item, the greater will be the activation of the hidden node that represents the exemplar.

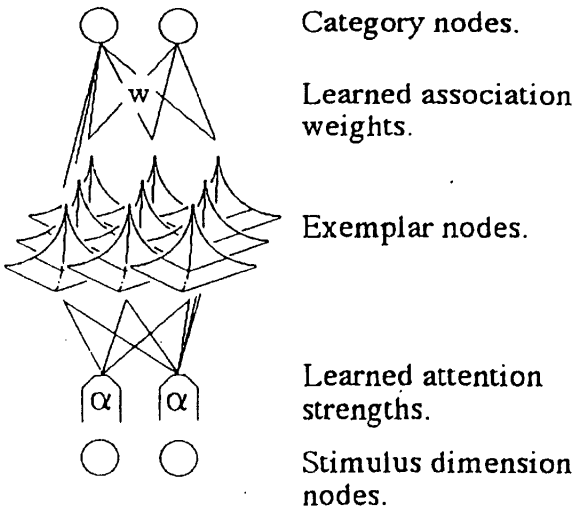

Figure 4. Illustration of the structure of the ALCOVE model.

The output of category node A is given by

$$O_A = \sum a_{ij}w_{jA}, \quad (2)$$

where w_{jA} is the (learned) association weight between exemplar node j and output node A, and the sum is over all exemplar nodes. The probability that item i is classified in category A is given by

$$P(A|i) = (O_A+b)/(O_A+O_B+2b), \quad (3)$$

where b is a background noise constant (Estes, 1994; Nosofsky & Kruschke, 1992; Nosofsky et al., 1992).

ALCOVE learns the attention weights, α_m, and association weights, w_{jA}, on a trial-by-trial basis by means of back propagation (Rumelhart et al., 1986). The precise learning rules are derived and presented by Kruschke (1992).

In the present application of ALCOVE, there are four free parameters: the overall sensitivity parameter (x) for scaling distances in the space (Equation 1); the background noise constant (b) for transforming category outputs into response probabilities (Equation 3); and learning rates, λ_w and λ_α, for updating the exemplar category association weights and dimensional attention weights, respectively (see Kruschke, 1992).

In a previous theoretical analysis, Nosofsky (1984) showed that the exemplar-based context model correctly predicted the order of difficulty for the six types of problems in Shepard et al.'s (1961) study, as long as it was assumed that subjects came to optimize the distribution of attention weights over the dimensions that compose the exemplars. The ALCOVE model, which extends the context model in a connectionist framework, provides a mechanism by which these optimal weights can be learned, and Kruschke (1992) verified that ALCOVE indeed predicts the correct ordering. The model learns to attend to the single relevant dimension that defines the Type I problem, to split attention between the two relevant dimensions that define the Type II problem, and to distribute attention optimally among the three dimensions that are relevant for solving the Types III–VI problems. The present research represents the first attempt, however, to test ALCOVE's ability to quantitatively predict the classification learning data in Shepard et al.'s paradigm.

Rational Model

According to Anderson's (1991) rational model, exemplars are grouped into clusters during the learning process. At any point during the learning sequence, the prob-

ability that an exemplar joins a cluster is determined jointly by the prior probability of each cluster and by the similarity of that exemplar to the cluster's central tendency. The prior probability of joining a cluster is determined by the current size of each cluster and by the value of a coupling parameter, which is a free parameter in the model. When the value of the coupling parameter is large, exemplars tend to join preexisting clusters, whereas when the value of the coupling parameter is small, exemplars tend to start their own new clusters. The similarity of an exemplar to a cluster's central tendency is computed by using an interdimensional multiplicative rule that is isomorphic to the similarity rule assumed in the context model (Anderson, 1990; Nosofsky, 1991).

There is a mechanism in the rational model for computing the probability that membership in each cluster signals a given category label. Thus, when an item is presented, the probability of its category label can be computed by summing the probability that it belongs to each cluster, weighted by the probability that each cluster signals a given category label.

In the present situation, the baseline version of the rational model has three free parameters: a coupling parameter (c) that influences the prior probability for items to join preexisting clusters or form new ones; a dimensional salience parameter (s_D) for computing similarities between items and the central tendencies of the clusters; and a category label salience parameter (s_L). In Anderson's framework, the category labels assigned to stimuli are treated in the same manner as values on the other stimulus dimensions. However, because the category label dimension has a special psychological status, it is reasonable to use a separate free parameter (s_L) to represent that the salience of the category label may be distinct from the saliences of the other stimulus dimensions. In addition to influencing the similarity between items and clusters, the category label salience parameter influences the estimated probabilities that individual clusters signal alternative category labels (for details, see Anderson, 1990, pp. 103, 105, 143).

To give the rational model additional flexibility for fitting data, we also provide it with a response-mapping parameter (r) for transforming estimated category label probabilities to observed response probabilities. Let p_A denote the rational model's estimated probability that item i signals category label A. Then the actual probability that the subject makes response A is given by

$$P_A = p_A^r/(p_A^r+p_B^r). \quad (4)$$

When $r = 1$, observed response probabilities match those that are estimated by the rational model, whereas when $r > 1$, observed response probabilities will be more extreme (closer to 0 or 1) than the estimated probabilities. (See Maddox & Ashby, 1993, for a previously successful use of such a response-mapping parameter in tests involving Nosofsky's, 1986, generalized context model.)

It is critical to understand that providing the rational model with this response-mapping parameter generalizes Anderson's (1990, 1991) presentation of the model. If $r = 1$, then the present model reduces to the version presented by Anderson (1990). Thus, this generalized model must fit the present classification learning data at least as well as the version without the response-mapping parameter. It is important to acknowledge that Anderson (1990, 1991) assumed only a monotonic relation between the estimated probabilities of the rational model and observed response probabilities. Nevertheless, it is critical to explore the ability of the model to yield good quantitative fits to data as well. Use of the present response-mapping parameter is a reasonable way to begin such an exploration.

In a previous theoretical analysis, Anderson (1991) showed that the rational model yielded fairly good qualitative predictions of the order of difficulty of learning the six problem types. Roughly, the model learns to classify items more efficiently when exemplars with the same category label are grouped into the same internal clusters, and exemplars with different category labels are grouped into distinct clusters. The Type I problem is learned most efficiently because all members of Category A are grouped into one cluster, and all members of Category B are grouped into a second cluster. The Type II problem is also learned efficiently, because only two distinct clusters are formed for the members of each category (e.g., Pairs 1-2 and 7-8 form clusters for category A—see Figure 1). Types III, IV, and V are learned less efficiently, because, depending on the precise sequence of learning exemplars, there is often a "singleton" cluster that is formed which consists of only one exemplar, and this last exemplar takes a long time to learn. Type VI is learned least efficiently because the categories break up entirely into singleton clusters.

Although promising in the respects noted above, the qualitative predictions of the rational model presented by Anderson (1991, Figure 13) also differed in certain ways from the data observed in the present study. For example, in contrast to the present results, the advantage for Type II over Types III, IV, and V did not emerge until around Block 5. Also, during the later learning blocks, performance on Type IV actually became worse than performance on Type VI. Finally, the overall level of performance on all of the problem types was considerably worse than that observed in the present study. It remains to be seen whether or not a more exhaustive search of the parameter space will locate parameters that allow the rational model to provide adequate quantitative fits to the classification learning data in Shepard et al.'s (1961) paradigm.

Configural-Cue Model

The configural-cue model (Gluck, 1991; Gluck & Bower, 1988b; Gluck et al., 1989), based on Rescorla and Wagner's (1972) description of classical conditioning, is a two-layered network model in which the connection weights are learned by the least mean squares (LMS) rule (Widrow & Hoff, 1960). The structure of the configural-cue network is illustrated in Figure 5. The in-

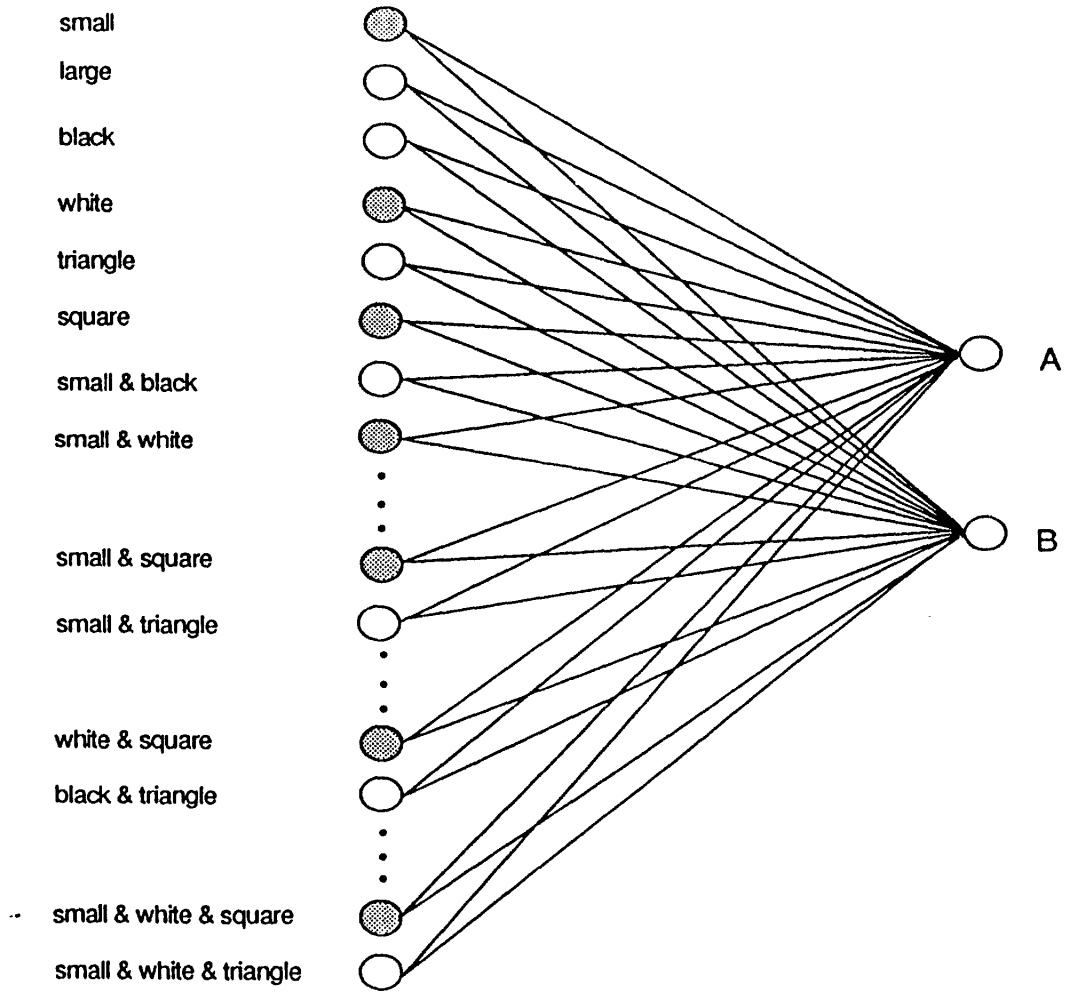

Figure 5. Illustration of the structure of the configural-cue model.

put nodes in the network code all configurations of features (i.e., single features, pairs of features, triples of features, etc.) that compose the items. Presentation of an n-dimensional stimulus pattern corresponds to presentation of the complete power set of all possible subsets of that pattern. For example, presentation of a small white square would cause seven input nodes to become active: small, white, square, small-white, small-square, white-square, and small-white-square (see Figure 5). In the present case, the stimuli are composed of three binary-valued dimensions, so the network has 26 input nodes (6 single nodes, 12 double nodes, and 8 triple nodes), 7 of which are activated on each trial. If input configuration i is present in a stimulus pattern on trial t, input node i receives an activation of 1, $a_i(t) = 1$; otherwise, the input node receives an activation of zero, $a_i(t) = 0$. Gluck (1991) demonstrated that this configural representation implies approximately the same exponential-decay stimulus generalization gradient found in ALCOVE and the rational model.

The activations on all input nodes are multiplied by the connection weights currently existing in the network, which are then summed to form outputs. Specifically, the output received by category node A is given by

$$O_A(t) = \sum a_i(t) w_{iA}(t), \tag{5}$$

where $w_{iA}(t)$ is the weight on the connection that links input node i to output node A on trial t. An analogous expression is used to compute the output received by category node B. The probability of a category A response is given by the same decision rule as that used in ALCOVE (Equation 3).

All weights in the network are initialized to zero. Learning of weights takes place by using the delta or least mean squares (LMS) rule (Widrow & Hoff, 1960). When cat-

egory A feedback is presented, output node A receives a teaching signal of $z_A(t) = 1$ and output node B receives a teaching signal of $z_B(t) = 0$. (When category B feedback is presented, the reverse teaching signals are provided.) The error at output node A is given by

$$\delta_A(t) = z_A(t) - O_A(t), \quad (6)$$

and likewise for the error at output node B. All weights in the network are then updated by using the rule

$$w_{iA}(t+1) = w_{iA}(t) + \lambda\delta_A(t)a_i(t), \quad (7)$$

where $w_{iA}(t)$ is the connection weight from input node i to output node A on trial t, and λ is the learning rate parameter. In our present tests of the configural-cue model, separate learning-rate parameters are allowed for the connections from single nodes, double nodes, and triple nodes to the output nodes. The standard configural-cue model thus has four free parameters: the background noise constant (b) in Equation 3, and learning rates λ_1, λ_2, and λ_3 for single, double, and triple nodes, respectively.[4]

Gluck and Bower (1988b) used the configural-cue model to simulate learning of the six Shepard et al. (1961) tasks. A single learning rate was used for all weights. Regardless of the learning rate, the model predicted a consistent ordering of difficulty. Most important, the overall ordering late in learning was found to be I < III < II < IV < V < VI. With the exception of misordering Types II and III, therefore, the configural-cue model was fairly successful at accounting for the relative difficulty of the six problems. As Gluck and Bower explain (1988b, pp. 187–189), the successes of the model derive from a combination of its configural representation of the stimuli, and the interactive, error-driven nature of the LMS learning rule, which learns associations for frequent and consistent cues more quickly than for infrequent and inconsistent cues.

Why, however, does the model fail to predict that people learn the Type II problem more quickly than Type III? Consider the structure of the Type II and Type III problems shown in Figure 1. There are no single-feature cues that are consistently associated with either category for either type of problem. For the Type II problem, there are only two double-feature cues that are consistently associated with category A (black-square and white-triangle), whereas for the Type III problem, there are three double-feature cues that are consistently associated with category A (small-black, large-square, and black-square). (An analogous set of cues with the same structure is associated with category B.) Because more cue configurations exist for classifying items in the Type III structure than in the Type II structure, the configural-cue model tends to learn the Type III structure more quickly.

Why do human subjects tend to learn the Type II structure more quickly? A likely reason is that in the Type II structure, both double-feature cues are defined over the *same dimensions* (i.e., color and shape for the example in Figure 1). By contrast, in the Type III structure, each of the relevant double-feature cues is defined over a different pair of dimensions (i.e., color-shape, size-shape, and size-color). Presumably, human subjects learn to attend selectively to *dimensions* when solving the classification problems, and this process facilitates the learning of the Type II problem. We hypothesized, therefore, that to improve the configural-cue model's ability to predict classification learning, mechanisms needed to be added that would allow the model to learn to selectively attend to dimensions.

Configural-Cue Model with Dimensionalized, Adaptive Learning Rates

We present here an extended version of the configural-cue model that we refer to as the *dimensionalized adaptive-learning rate* (DALR) model (Gluck, Glauthier, & Sutton, 1992). This DALR model is based on recent advances in adaptive-learning procedures for neural networks (Jacobs, 1988; Sutton, 1992a, 1992b). Most important for the present purposes, the DALR model incorporates mechanisms that allow the configural-cue model to learn to selectively attend to dimensions.

Two key ideas underlie the DALR model. First, whereas in the standard configural-cue model the learning rates on each of the connections remain constant during training, in the extended model there is a meta-learning process that dynamically modifies each learning rate. Incorporating a dynamically modifiable learning rate provides a mechanism for adjusting the salience of individual cues. Specifically, the model learns to assign higher learning rates to those connections in the network that are highly diagnostic for solving a given problem. From a psychological perspective, this enhanced learning rule is analogous to adding an attentional component to the configural-cue model. Attention is conceptualized here as an enhanced tendency to change the strength of a given cue's associations. (See Frey & Sears, 1978; Mackintosh, 1975; and Pearce & Hall, 1980, for similar psychological ideas about the learning of cue-specific saliences and associabilities.)

A second key idea behind the DALR model is that it is provided with a "knowledge" of the underlying dimensional structure of the stimuli. In the standard configural-cue model, information about the dimensional structure of the stimuli is basically discarded from the input representation. For example, the model does not represent the relationship between *black* and *white* as being any different from the relationship between *black* and *large*. Each dimension value is represented by a unique input node that is either activated or not activated, depending on its presence or absence in the input pattern.

It follows that even with the metalearning process described above, there is still no form of *dimensionalized* attention learning in the standard configural-cue model, because separate learning rates can arise for each distinct configuration of dimension values. For example, separate learning rates can occur for the input nodes black-square and white-triangle, which are defined over the same dimensions.

Thus, we introduce a structural modification to the model by *linking* the learning rates for given dimensions and configurations of dimensions. For example, connec-

tions from the nodes for "large" and "small" share a single learning rate, allowing attention (in the form of a high learning rate) to be focused on the entire dimension of "size" rather than on specific dimension values. When the cue "small" is relevant to the classification task, the network increases the linked learning rate for size, thus assigning more attention to the cue "large" as well. Likewise, all connections from double nodes defined over the dimensions of size and shape (large-square, large-triangle, small-square, small-triangle) share a single learning rate, and so forth.

We describe the formal implementation of this extended DALR model in the Appendix. We use five free parameters to fit the DALR model to the classification learning data: an initial learning rate (λ_0) that applies to all connections in the network; a background noise constant (b) that is used in the decision rule (Equation 3); and three separate metalearning parameters for adapting the learning rates on connections from single, double and triple nodes (θ_1, θ_2, and θ_3, respectively). These metalearning parameters are used to dynamically adjust the learning rates on each of the (dimensionalized) connections in the network.

FITTING THE MODELS TO THE LEARNING DATA

Because there were essentially zero errors for most of the problems past Block 16, we fitted the models to the learning data from Blocks 1–16 only. The learning curves were fitted with all parameters held fixed across the six problems. We used two criteria of fit for evaluating the models. The first criterion was to minimize the sum of squared deviations (*SSD*) between predicted and observed error proportions across all 6 problems and 16 learning blocks (96 data points). A difficulty with using *SSD* as a criterion of fit, however, is that not all data points have the same error variance. Proportions close to zero or unity have very small error variances, so deviations in these regions are potentially quite important. Thus, we also fitted the models by using a weighted least squares criterion. The weighted sum of squared deviations (*WSSD*) is found by summing the squared deviations between predicted and observed error proportions weighted by $1/\sigma^2$, the inverse of the variance of each cell proportion (see, e.g., Bishop, Fienberg, & Holland, 1975). Assuming binomial variability, the error proportion q_i has variance $q_i(1-q_i)/N$, where N is the number of observations.[5] In this article, we report fits based on both the unweighted and weighted *SSD* measures. Fortunately, in our analyses of the main data set, both measures lead to the same conclusions, and we focus primarily on the unweighted *SSD* measure in discussing the results.

To fit each of the models, 100 random stimulus sequences were generated. The characteristics of each random sequence matched the constraints in our experimental design. For any given set of parameters, a model was used to generate predictions based on each random sequence. These 100 sets of predicted values were then averaged, and the averaged values constituted the predictions that were fitted to the observed data. The ALCOVE, configural-cue, and DALR models were fitted to the data by using a hill-climbing parameter search routine. Various different starting configurations of the parameter values were used in these searches, in an attempt to guard against local minima. The same hill-climbing routine was also used to fit the rational model. However, the rational model may be especially prone to local minima, because even small changes in parameter values can sometimes lead to dramatically different predictions from that model. (Small parameter changes can lead the model to form different clusters, resulting in markedly different behavioral predictions.) In a special attempt to guard against local minima for the rational model, therefore, a "grid search" was also used, which consisted of checking a huge number of possible combinations of parameter values. Both the hill-climbing and the grid search methods led to the same best-fitting parameters.

MODEL-FITTING RESULTS

The summary fits for ALCOVE, the standard configural-cue model, the DALR model, and the rational model are reported in Table 2, and the best-fitting parameters are reported in Table 3. The predictions of the four models, based on the minimum *SSD* criterion, are shown graphically in Figures 6–9.

Among the four models, ALCOVE provides the best overall quantitative account of the learning data. ALCOVE accounts for 95.5% of the variance in the learning data for the six problems taken together. The remaining three models, which provide roughly the same overall fits as one another, account for an average of 86.3% of the variance. The advantage for ALCOVE is fairly consistent across all the problem types, as indicated by the individual-problem *SSD*s that are reported in Table 2. Furthermore, the advantage holds regardless of whether one adopts the *SSD* or *WSSD* criterion of fit (see Table 2).

Visual comparison of Figures 6 and 2 reveals that ALCOVE does quite well at capturing the pattern of learning data. It predicts a clear advantage for the Type I problem, followed by Type II, by Types III, IV, and V, which are roughly equivalent, and finally by Type VI, which lags clearly behind.

By comparison, inspection of Figures 7–9 reveals obvious shortcomings in the predictions of the remaining models. As expected, the standard configural-cue model (Figure 7) predicts incorrectly that the Type III problem is learned more quickly than the Type II problem. Furthermore, the model predicts only a slight advantage for Type II over Types IV and V, in contrast to the observed data. There is also not enough of an advantage for the Type I problem over the remaining problem types.

The extended DALR configural-cue model (Figure 8) improves on the standard model, at least with regard to the qualitative trends noted above—but the improvement

Table 2A
Sum of Squared Deviations (*SSD*) Between Predicted and Observed Error Probabilities in Blocks 1–16 for Problems I–VI

Model	Problem I	II	III	IV	V	VI	*SSD*	*RMSD*	%Var
ALCOVE	.002	.012	.004	.005	.011	.028	.061	.025	95.5
Configural cue	.025	.034	.036	.024	.012	.039	.172	.042	87.5
DALR	.021	.032	.039	.043	.005	.071	.211	.047	84.6
Rational	.061	.026	.026	.025	.011	.033	.182	.044	86.8

Note—*RMSD* = root mean squared deviation. %Var = percentage of variance accounted for. All models were fitted by searching for the fixed set of parameters that minimized overall *SSD* across the six problem types. The resulting *SSD*s for each individual problem type are also shown.

Table 2B
Weighted Sum of Squared Deviations (*WSSD*) Between Predicted and Observed Error Probabilities in Blocks 1–16 for Problems I–VI

Model	Problem I	II	III	IV	V	VI	*WSSD*
ALCOVE	.024	.306	.028	.112	.104	.297	0.871
Configural cue	.214	.285	.519	.502	.113	.431	2.063
DALR	.087	.360	.402	.532	.078	.422	1.881
Rational	.737	.123	.719	.821	.630	1.515	4.545

Note—All models were fitted by searching for the fixed set of parameters that minimized overall *WSSD* across the six problem types. The resulting *WSSD*s for each individual problem type are also shown.

is not sufficient to yield a satisfactory quantitative fit. By incorporating the dimensionalized adaptive learning rates, the extended model correctly predicts the rapid learning that occurred for the Type I problem. In addition, it now predicts slightly better performance on the Type II problem than on the Type III problem. But the quantitative difference in performance between Types II and III is far too small in magnitude in relation to what is observed in the data. We also tested versions of the DALR model in which separate initial learning rates were allowed for connections from single, double, and triple nodes, but these elaborated models provided only modest improvements in quantitative fit.

An intriguing question is why the DALR model fails to predict enough of an advantage for the Type II problem over Type III, despite the fact that a dimensionalized selective-attention mechanism has been added to that model. In an attempt to answer this question, we tracked the connection weights that were learned by the best-fitting version of the DALR model. As expected, for the Type II problem, the model learned to assign very large weights to the two perfectly diagnostic doublet cues for each category. For example, if we use the illustration from Figure 1, the model learned to assign large weights to the connections linking black-square and white-triangle to category *A*. No other connections received very much weight. Also as expected, for the Type III problem, the model learned to assign large weights to the three perfectly diagnostic doublet cues for each category (black-square, large-square, and small-black for category *A*). Because these doublet cues are defined over different pairs of dimensions, however, the weights assigned to them were not as large as those for the Type II problem. Thus, this aspect of the DALR model worked precisely as anticipated.

Table 3
Best-Fitting Model Parameters

Model	Parameters
	SSD Criterion
ALCOVE	$x = 6.330$, $\lambda_w = .179$, $\lambda_\alpha = .409$, $b = .011$
Configural cue	$\lambda_1 = .000$, $\lambda_2 = .048$, $\lambda_3 = .079$, $b = .015$
DALR	$\lambda_0 = .064$, $\theta_1 = .104$, $\theta_2 = .463$, $\theta_3 = .010$, $b = .028$
Rational	$c = .318$, $s_D = .488$, $s_L = .046$, $r = .930$
	WSSD Criterion
ALCOVE	$x = 5.092$, $\lambda_w = .246$, $\lambda_\alpha = .442$, $b = .00092$
Configural cue	$\lambda_1 = .046$, $\lambda_2 = .082$, $\lambda_3 = .065$, $b = .0011$
DALR	$\lambda_0 = .059$, $\theta_1 = .377$, $\theta_2 = .175$, $\theta_3 = .000$, $b = .0017$
Rational	$c = .318$, $s_D = .488$, $s_L = .046$, $r = 1.496$

Note—*SSD*, sum of squared deviations; *WSSD*, weighted sum of squared deviations.

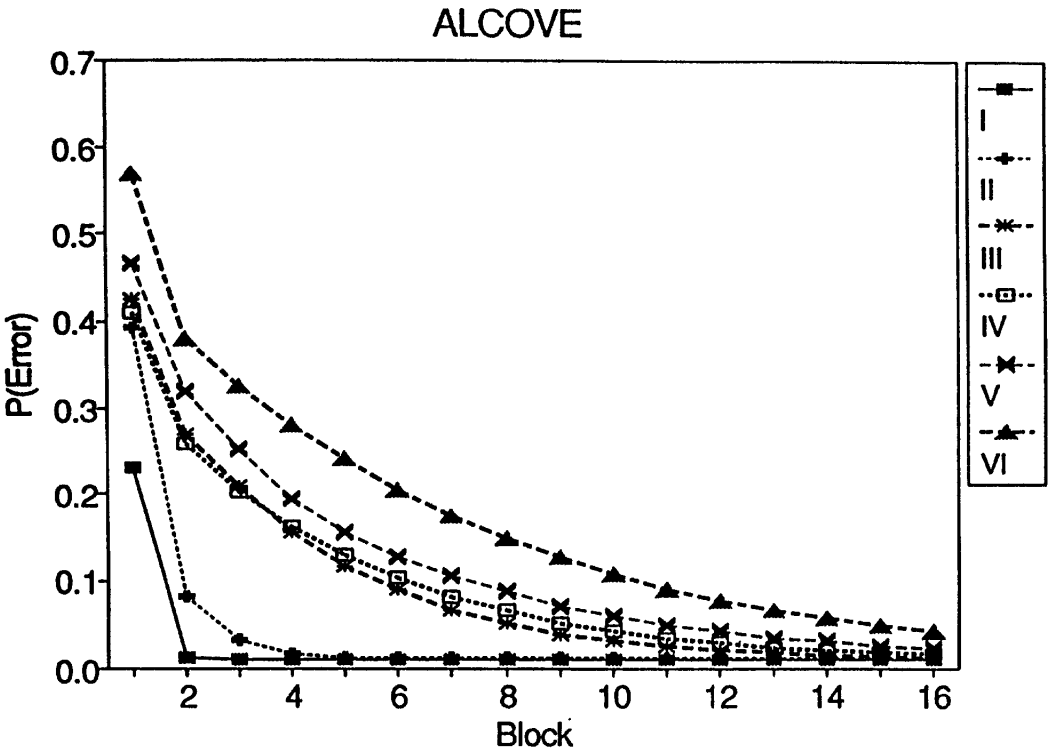

Figure 6. Learning curves for the six problem types predicted by ALCOVE.

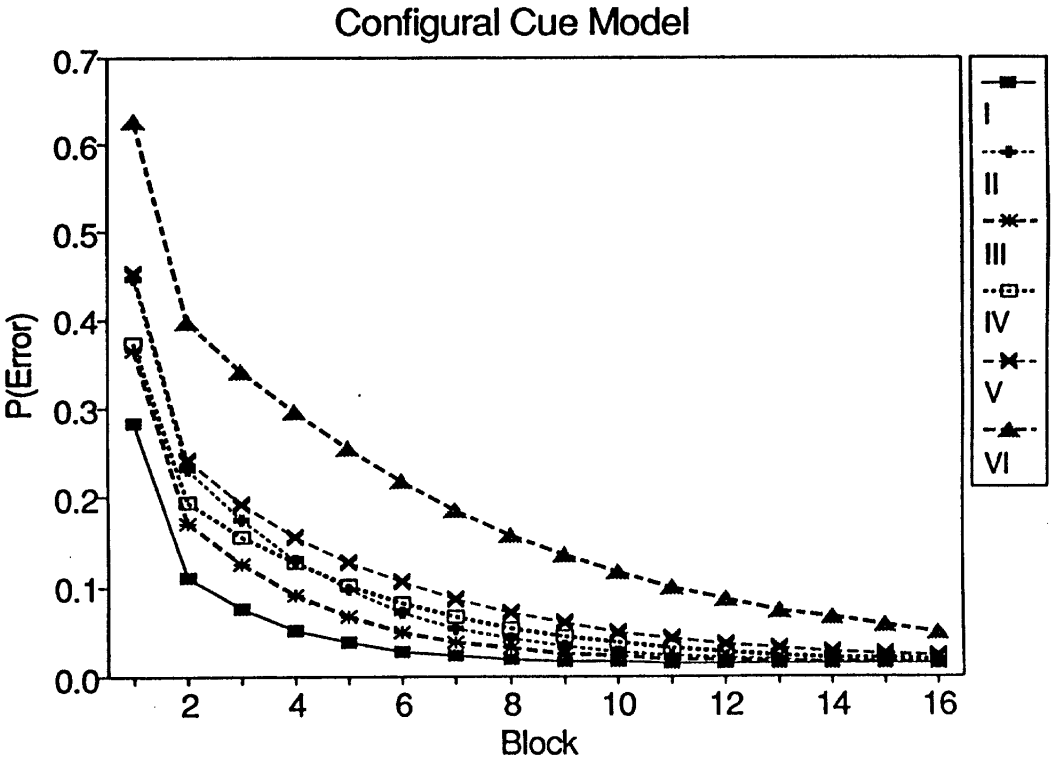

Figure 7. Learning curves for the six problem types predicted by the configural-cue model.

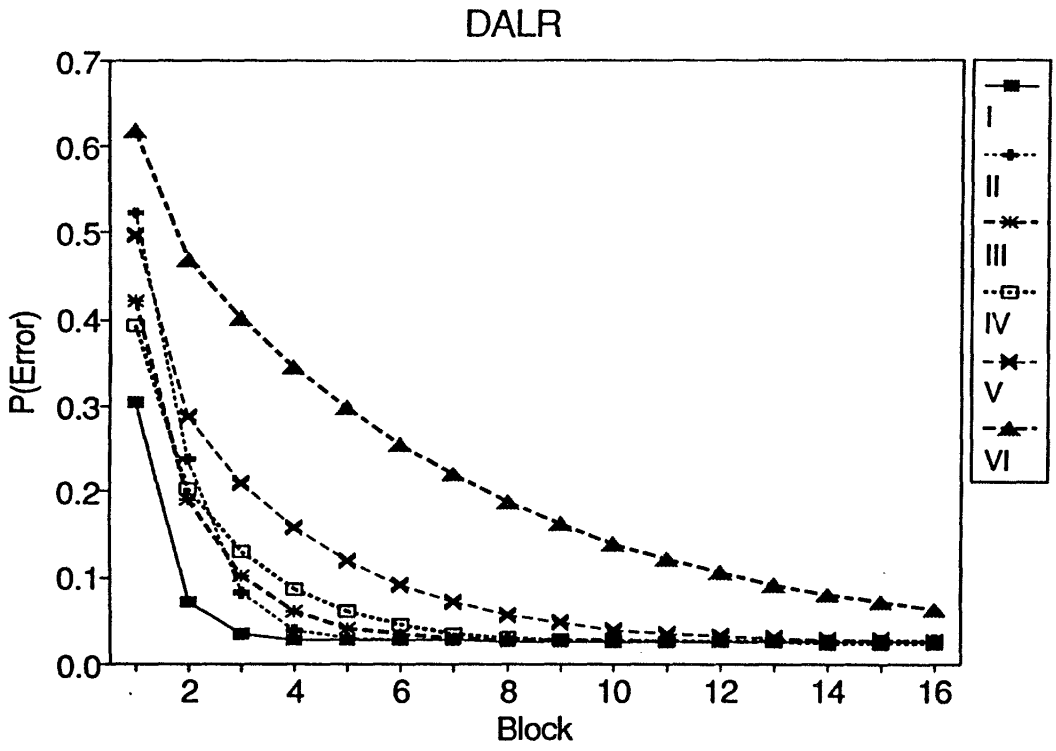

Figure 8. Learning curves for the six problem types predicted by the DALR model.

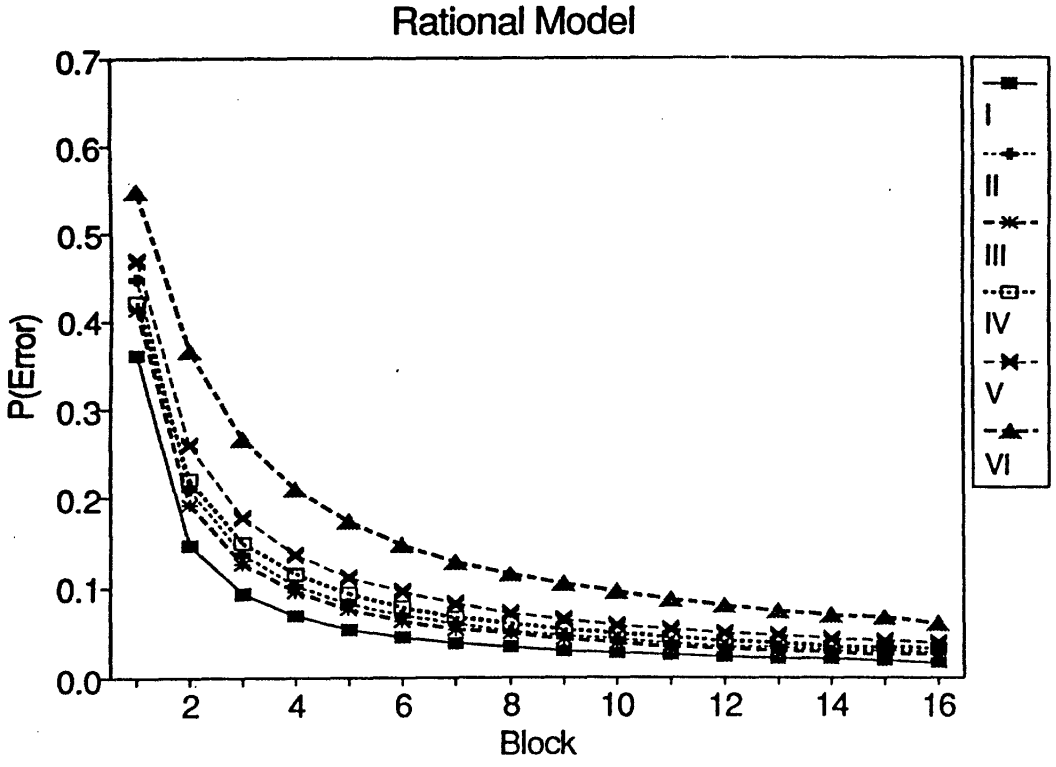

Figure 9. Learning curves for the six problem types predicted by the rational model.

A surprising result, however, was that for the Type III problem, the DALR model also learned to assign moderately large weights to connections from some of the singlet cues. Consider again Figure 1, and note that for the Type III problem, the *individual* dimensions of shape and color, although not perfectly diagnostic, are partially diagnostic of category membership. Three of the four squares are assigned to category *A*, and three of the four triangles are assigned to category *B*. Likewise, three of the four black objects are assigned to category *A*, and three of the four white objects are assigned to category *B*. Thus, overall performance is benefited if, in addition to attending to the perfectly diagnostic doublet cues, the model learns to attend to the partially diagnostic singlet cues. Indeed, the nature of the adaptive-learning-rate mechanism in the DALR model is to assign higher learning rates to whatever cues are diagnostic of category membership. In a nutshell, then, a plausible explanation for why the DALR model learns the Type III problem nearly as efficiently as the Type II problem is because it takes simultaneous advantage of both the doublet-cue associations and singlet-cue associations that are relevant to solving that problem. We provide further discussion of the implications of this learning process in our General Discussion.

Like the configural-cue model, the rational model (Figure 9) also predicts better performance on the Type III problem than on the Type II problem, at least with the present parameter settings. Another shortcoming of the rational model is that learning on the Type I problem occurs too slowly, and there is not much separation between the Type I learning curve and the remaining learning curves. As we noted earlier, Anderson (1991, Figure 13) reported predictions for the rational model in which different parameter settings were used, and in which performance was markedly better on the Type I and Type II problems than on the remaining problems. Unfortunately, with these alternative parameter settings, the rational model vastly underpredicts the overall level of performance that was observed in our experiment. (It also mispredicts other qualitative features of the learning data that we discussed earlier.) We have been unable to find parameter settings for the rational model that yield performance levels close to those observed in our experiment while at the same time correctly ordering the difficulty of the problem types. It may be that, like the configural-cue model, a form of dimensionalized attention learning needs to be added to the rational model to improve its quantitative predictions. In addition, perhaps an alternative mechanism for converting internal estimated probabilities to observed response probabilities would yield improved quantitative fits.

The attention-learning mechanism in ALCOVE is critical in allowing that model to correctly predict the data. A restricted version of ALCOVE arises by holding fixed the attentional learning rate at $\lambda_\alpha = 0$. This restricted model yields a fit that is far worse than that of the full model (*SSD* = .202, *RMSD* = .046, %variance = 85.3). It predicts worse performance on the Type II problem than on Types III and IV, and performance on Type II that is nearly as poor as that on Type V. Learning on the Type I problem also proceeds too slowly. This conclusion about the importance of attention learning in exemplar-based models of classification agrees with earlier ones reached by Nosofsky (1984) and Kruschke (1992).

GENERAL DISCUSSION

The main purpose of this research was to collect a rich set of learning data that would be useful for providing quantitative tests of current models of classification learning. We used Shepard et al.'s (1961) classic paradigm, but we collected more detailed data and tested more subjects than in the original study. Our general patterns of results replicated those observed in the original study. However, by recording detailed learning curves for each of the problems, we obtained a data set suitable for quantitative testing.

Our initial quantitative tests were focused on ALCOVE, the rational model, and the configural-cue model. These models are among the leading candidates in the field today, but there have been few attempts to develop quantitative comparisons to test among them. When learning data from all six problem types were analyzed simultaneously, ALCOVE provided the best overall fit, yielding learning curves that matched the observed data nicely. A clear shortcoming of the configural-cue model and the rational model is that they failed to predict the magnitude of the advantage observed for the Type I and Type II problems over Types III–V and VI.

Because only a subset of dimensions is relevant for solving the Types I and II problems, it seems likely that some form of dimensionalized, selective attention process is involved, and this process is well captured by the attention-learning mechanism in ALCOVE. It seems likely that the quantitative predictions of the configural-cue model and the rational model could be improved if they too incorporated forms of dimensionalized attention learning.

We made an attempt to incorporate a dimensionalized attention-learning mechanism in the elaborated DALR configural-cue model. Although this mechanism led to improvements in the qualitative predictions of the model, it still had quantitative shortcomings. An analysis of the pattern of connection weight learning in that model indicated that a likely reason for its quantitative shortcomings is that the model learned to attend simultaneously to multiple configurations of cues (e.g., black-square, large-square, small-black, square, black). In other words, it attempted to take simultaneous advantage of all and whatever forms of diagnostic evidence were available in the stimulus patterns. It may be that people are more naturally inclined to learn to attend selectively to a more limited set of cues. If this suggestion is correct, further improvements in the configural-cue model could probably be achieved by designing certain capacity limits into the attention-learning mechanisms. We leave this idea as a project for future exploration and research.

Another issue is that the present models were fitted to averaged learning data, and not to the data of individual subjects. Although averaged data may sometimes obscure patterns observed at the individual subject level, we have good reason to believe that, in the present study, the averaged data are at least fairly reflective of individual subject behavior. Before generating our averaged learning curves, we inspected histograms of individual subject errors on each of the problem types. These histograms appeared to be symmetric and unimodal, with no extreme outliers. We believe that these histograms of individual subject errors could be fitted quite well by simply allowing variability in the parameter values (e.g., the learning rates) across individual subjects. However, we leave this ambitious goal of actually fitting distributions of individual subject behaviors as another issue for future research.

The learning tasks used in this study involved rule-based categories. In another recent theoretical investigation involving rule-based categories, Choi, McDaniel, and Busemeyer (1993) compared alternative network models on their ability to predict patterns of classification learning data in some concept identification tasks reported by Salatas and Bourne (1974). Choi et al. (1993) found that as long as one incorporated certain prior biases into the structure of the network, ALCOVE again provided an excellent description of the pattern of results. Thus, the applicability of ALCOVE in this domain of learning rule-based categories appears to have some generality. Whether similar patterns of results will be observed in situations involving more ill-defined categories remains an open question. Conceivably, alternative classification strategies operate under different learning conditions, and the configural-cue model and the rational model may show advantages in other domains.

A central claim of Shepard et al. (1961) was that their results implicate some form of abstraction or selective attention to dimensions during the course of classification learning. This attention-learning process is in some sense limited in capacity. Focusing attention on fewer dimensions allows for better performance on classifications that are defined over those dimensions. This same basic message echoes loudly in the present replication and extension of their study. Models of stimulus generalization and cue conditioning in human learning that do not incorporate some form of limited-capacity, selective attention mechanism for dimensionalized stimuli have difficulty accounting for the pattern of results observed in these fundamental classification tasks.

REFERENCES

ANDERSON, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

BISHOP, Y. M. M., FIENBERG, S. E., & HOLLAND, R. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

CHOI, S., MCDANIEL, M. A., & BUSEMEYER, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition*, **21**, 413-423.

ESTES, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, **115**, 155-174.

ESTES, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

ESTES, W. K., CAMPBELL, J. A., HATSOPOULOS, N., & HURWITZ, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 556-571.

FREY, P. W., & SEARS, R. J. (1978). Models of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review*, **85**, 321-340.

GLUCK, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, **2**, 50-55.

GLUCK, M. A., & BOWER, G. H. (1988a). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227-247.

GLUCK, M. A., & BOWER, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory & Language*, **27**, 166-195.

GLUCK, M. A., BOWER, G. H., & HEE, M. (1989). A configural-cue network model of animal and human associative learning. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society* (pp. 323-332). Hillsdale, NJ: Erlbaum.

GLUCK, M. A., GLAUTHIER, P. T., & SUTTON, R. S. (1992). Adaptation of cue-specific learning rates in network models of human category learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society* (pp. 540-545). Hillsdale, NJ: Erlbaum.

JACOBS, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, **1**, 295-307.

KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

MACKINTOSH, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, **82**, 276-298.

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

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

NOSOFSKY, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.

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

NOSOFSKY, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, **2**, 416-421.

NOSOFSKY, R. M., & KRUSCHKE, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. *Psychology of Learning & Motivation*, **28**, 207-250.

NOSOFSKY, R. M., KRUSCHKE, J. K., & MCKINLEY, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 211-233.

PEARCE, J. M., & HALL, A. G. (1980). A model for Pavlovian conditioning: Variation in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, **87**, 532-552.

RESCORLA, R. A., & WAGNER, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol 1. Foundations* (pp. 318-362). Cambridge, MA: Bradford Books/MIT Press.

SALATAS, H., & BOURNE, L. E., JR. (1974). Learning conceptual rules:

III. Processes contributing to rule difficulty. *Memory & Cognition*, 2, 549-553.

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

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

SUTTON, R. S. (1992a). Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 171-176). Cambridge, MA: MIT/AAAI Press.

SUTTON, R. S. (1992b). Gain adaptation beats least squares? In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems* (pp. 161-166). Yale University, Center for Systems Science.

WIDROW, G., & HOFF, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, 4, 96-194.

NOTES

1. By recoding dimension values, Type VI becomes the parity problem. Assign logical value 1 to dimension values square, black, and large, and assign value 0 to dimension values triangle, white, and small. Any stimulus with an odd number of 1s belongs to Category 1, whereas any stimulus with an even number of 1s belongs to Category 2.

2. Technically, the t tests reported for the pooled data are invalid, because each subject participated in two problems and the observations across groups are therefore not completely independent. Because the same statistical results were obtained for the first problem taken by itself, however, this does not appear to be a major concern with respect to describing the general pattern of results.

3. The learning curves for the individual item types and the detailed results of fitting the alternative models to the item-type data are available from the authors on request.

4. Some additional technical details regarding the configural-cue model and ALCOVE include the following. First, on rare occasions, the outputs for the configural-cue model can take on slightly negative values. When the Equation 3 response rule is applied, negative outputs are truncated to zero. Second, in calculating the error signal (Equation 6), we found that "humble" teachers provided slightly better fits than "strict" teachers did. (The same finding was observed for ALCOVE.) When a humble teacher is used, the error is defined to be zero if the magnitude of the output, in either the positive or the negative direction, exceeds the magnitude of the teaching signal (see Kruschke, 1992, for details). The results reported in this article are for the models that used humble teachers, but the general conclusions do not depend on which type of teaching signal is assumed. Finally, we also fitted versions of ALCOVE and the configural-cue model that used exponential transformations of outputs in the choice rule, instead of raw outputs, but these versions gave slightly worse fits than the present versions did.

5. Because in the present experiment all cell proportions were based on the same number of observations, however, N was simply deleted from the *WSSD* computation. In addition, in cases in which there were zero error proportions, we adopted the convention that $q_i = 1/2N$ (see, e.g., Bishop et al., 1975).

APPENDIX
Formal Description of the DALR Model

In this Appendix, we provide a formal description of the manner in which the dimensionalized, adaptive learning rate (DALR) model extends the standard configural-cue model. A description of the basic (nondimensional) ALR network model of classification learning can be found in Gluck et al. (1992). As noted in the text, the DALR model is based on recent advances in adaptive-learning-rate methods for neural networks. These methods are motivated primarily by engineering considerations, with the goal of increasing the learning efficiency of the networks. It is an open question whether or not these methods have a natural psychological or neural interpretation.

Adaptive-learning-rate methods are metalearning algorithms for adapting step-size parameters (i.e., learning rates) during a learning process, which in the present case is the LMS rule. Consider, for example, one of the weights in a connectionist network and how it changes over time. If the weight changes are all in the same direction (e.g., all increases) this signifies that the step-size parameter is too small. The weight could reach its asymptotic value more quickly if it took larger steps. On the other hand, if the weight changes are in opposite directions (e.g., first up and then down) this signifies that the step-size parameter is too large. The basic idea behind current adaptive-learning-rate methods is to adjust the step size according to the correlation between successive weight changes, with the goal of obtaining zero correlation. (A positive correlation signifies that weight changes have been in the same direction; a negative correlation signifies that weight changes have been in opposite directions.) The particular mechanism used to form the dimensionalized adaptive-learning-rate configural-cue model tested in this article was proposed by Sutton (1992a) and is known as the *incremental delta-bar-delta* (IDBD) *algorithm*.

Specifically, with this IDBD algorithm, there is a different learning rate λ_{iA} for each connection from input node i to output node A, and similarly for connections to any other output nodes. These learning rates change according to a metalearning process. The standard learning rule (Equation 7) becomes

$$w_{iA}(t+1) = w_{iA}(t) + \lambda_{iA}(t+1)\delta_A(t)a_i(t), \qquad \text{(A1)}$$

where the learning-rate parameters are now indexed by trials, $t+1$. (The λ_{iA} are indexed by $t+1$ rather than t to indicate that their update, by a process described below, occurs before the w_{iA} update.) To ensure that the learning rates remain positive, they are expressed and stored in the form $\lambda_{iA}(t) = \exp[\beta_{iA}(t)]$. The IDBD algorithm updates the β_{iA} by

$$\beta_{iA}(t+1) = \beta_{iA}(t) + \theta\delta_A(t)a_i(t)h_{iA}(t)/\sqrt{\lambda_{iA}}\,(t), \qquad \text{(A2)}$$

where θ is a positive constant, the *metalearning rate*, and h_{iA} is a memory variable associated with each connection from input node i to output node A. The memory variable indicates whether the recent errors associated with a given connection weight have been positive or negative. To indicate this information, the memory variable is initialized at zero and updated by:

$$h_{iA}(t+1) = h_{iA}(t)[1-\lambda_{iA}(t+1)a_i^2(t)]^+ + \lambda_{iA}(t+1)\delta_A(t)a_i(t), \qquad \text{(A3)}$$

where $[x]^+$ is x, if $x > 0$, else 0. The first term in the preceding equation is a decay term—the product $\lambda_{iA}(t+1)a_i^2(t)$ is normally zero or a positive fraction, and this causes a decay of h_{iA} toward zero. The second term increments h_{iA} by the previous error, assuming that cue i was present in the input pattern [$a_i(t) = 1$]. The memory, h_{iA}, is thus a decaying trace of the cumulative sum of recent errors associated with the presence of cue i (Sutton, 1992a).

Note from Equation A2 that if the present error, $\delta_A(t)$, matches the sign of the cumulative sum of recent errors, $h_{iA}(t)$, the learning rate on the connection from input node i to output node A will be increased. If the signs mismatch, the learning rate will be decreased. This algorithm, therefore, provides one way of instantiating the key idea behind adaptive-learning-rate methods—

namely, that of adjusting learning-rate parameters according to the correlation between successive weight changes.

As we discussed previously in the text, the DALR model introduces a form of *dimensionalized* attention learning by linking the learning rates for given dimensions and configurations of dimensions. As noted earlier, in the present experimental design, the configural-cue model has 26 input nodes. By linking learning rates, however, we reduce the number of unique learning rates from 26 to 7 for each of the two output nodes. (Unique learning rates occur on connections from nodes defined over the dimensions of size, shape, color, size-shape, size-color, shape-color, and size-shape-color.) Note that each connection retains its own unique history variable, h_{iA}; thus, there are 7 distinct learning rates and 26 history variables for each of the two outputs.

(Manuscript received April 16, 1993;
revision accepted for publication August 27, 1993.)

MEMORY & COGNITION

COMPARING MODELS OF RULE-BASED CLASSIFICATION R. NOSOFSKY, T. AMERICAN PSYCHOLOGICAL ASSN.

1994 REPRINTED WITH PERMISSION VIA THE COPYRIGHT CLEARANCE CENTER

PSYCHOLOGICAL REVIEW

RULE-PLUS-EXCEPTION MODEL OF CLASSIFICATION R, NOSOFSKY, T. AMERICAN PSYCHOLOGICAL ASSN

1994 REPRINTED WITH PERMISSION VIA THE COPYRIGHT CLEARANCE CENTER