Memory & Cognition
1993, 21 (4), 413-423

Incorporating prior biases in network models of conceptual rule learning

SANGSUP CHOI, MARK A. McDANIEL, and JEROME R. BUSEMEYER
Purdue University, West Lafayette, Indiana

A series of simulations is reported in which extant formal categorization models are applied to human rule-learning data (Salatas & Bourne, 1974). These data show that there are clear differences in the ease with which humans learn rules, with the conjunctive the easiest and the biconditional the hardest. The original ALCOVE model (an exemplar-based model), a configural-cue model, and two-layer backpropagation models did not fit the rule-learning data. ALCOVE successfully fit the data, however, when prior biases observed in human rule learning were implemented into weights of the network. Thus, current empirical learning models may not fare well in situations in which learners enter the concept-formation situation with preconceived biases regarding the kinds of concepts that are possible, but such biases might nevertheless be captured within these models. By incorporating preexperimental biases, ALCOVE may hold promise as a comprehensive category-learning model.

Theoretical and empirical work in human concept learning has undergone considerable transition since Hull's (1920) seminal master's thesis. Hull's work and the predominant stimulus-response theories of the time led to an initial focus on concepts that were defined in terms of specific stimulus attributes (e.g., size, shape) related through bidimensional logical rules such as conjunction, disjunction, and so on (cf. Horton & Turnage, 1976). By the mid-1970s, an impressive body of literature had accrued on concepts of this type, with extensive experimental and theoretical analysis devoted to the question of how humans learn these rule-based concepts (Bourne, 1967, 1974).

Very little attention has been devoted to rule learning in recent years, however, because of a shift in the kinds of concepts emphasized in current empirical and theoretical work. Research has been redirected to ill-defined (fuzzy) concepts that cannot be defined absolutely by a simple combination of certain attributes (see, e.g., Medin & Schaffer, 1978; Rosch, 1975), to "abstractionist" paradigms in which continuous stimulus dimensions not made explicit to the learner are used to construct a continuous range of stimuli (see, e.g., Homa, 1984; Posner & Keele, 1968), or to probabilistic categories in which the critical stimulus values are not deterministically related to category membership (see, e.g., Estes, Campbell, Hatsopoulus, & Hurwitz, 1989; Gluck & Bower, 1988a). Accompanying the shift in emphasis to ill-defined and nondeterministic concepts has been the popularization of new theoretical approaches, two of the most prominent being the exemplar frameworks of categorization (see, e.g., Brooks, 1978; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1987, 1988) and the adaptive network models of category learning (see, e.g., Estes et al., 1989; Gluck & Bower, 1988a; Shanks, 1991). These models have enjoyed enough success to stimulate considerable interest among concept-learning theorists. First, they provide formal, quantitatively based theoretical vehicles for accounting for and predicting categorization behavior, allowing more detailed and principled instantiation of models than earlier qualitatively based approaches allowed (cf. Homa, Sterling, & Trepel, 1981). Second, the hope (sometimes expressed in explicit claims; see, e.g., Nosofsky, 1991) is that these current models that have evolved from the consideration of less well-defined categories will serve as general accounts of human concept learning, irrespective of the nature of the concept.

When one considers the broad applicability of current concept-learning approaches, one must remember that these models emerged specifically to account for the findings and data associated with categorization behavior of ill-defined and nondeterministic categories. There have been few attempts to directly assess the applicability of these models to the wealth of findings concerning human learning of well-defined concepts (a few notable exceptions are the efforts to simulate the results of Shepard, Hovland, & Jenkins, 1961, with exemplar-based and adaptive network models; see Gluck & Bower, 1988b; Kruschke, 1992; Nosofsky, 1984). The absence of attempts in the literature to empirically validate computational models of rule learning is especially serious from the perspective of adaptive network models, because the

This work was supported by National Institute of Mental Health Grant MH 47126 to J.R.B. and M.A.M. We would like to thank Lyle Bourne for providing the unpublished manuscripts giving the details of the data and the procedures. We also thank Lee Brooks, William Estes, Mark Gluck, Rob Nosofsky, and an anonymous reviewer for their useful comments on an earlier version of the paper. Correspondence concerning this article should be addressed to M.A. McDaniel, Department of Psychological Sciences, Purdue University, 1364 Psychological Sciences Building, West Lafayette, IN 47907-1364 (e-mail: mcdaniel@psych.purdue.edu).

cornerstone of their resurrection (from the perceptrons of Minsky & Papert, 1969) was the demonstration that adaptive networks could learn logical conceptual rules such as the exclusive disjunctive (Rumelhart, Hinton, & Williams, 1986)—rules for which human learning is well documented (Bourne, 1974; Neisser & Weene, 1962).

Clearly, demonstrating sufficiency is only a first step in evaluating the fruitfulness of these models. Given the richness of the empirical literature on human conceptual rule learning, it is imperative that the currently prominent models be examined from the standpoint of how well their learning performance parallels that demonstrated by human learners.

More precisely, some of these models have proved to be well suited for learning the relevant dimensions that provide the basis for classifying a set of stimuli (see Kruschke, 1992). However, rule learning embodies another component of concept acquisition, that of learning the critical *relations* among the relevant attributes (this component of the concept might be viewed as the abstract part of the concept). Empirical work suggests that unique behaviors are associated with rule learning (Bourne & Guy, 1968), and some have warned that otherwise successful formal models (e.g., exemplar-based models) may not capture the processes involved in rule-learning (Kruschke, 1992).

This paper describes our attempts to apply the extant formal models of categorization to a prominent and well-replicated set of rule-learning data. For pragmatic reasons, we limit our report to the models that have been the most successful (and have attracted the most interest) in accounting for human concept learning, at least for paradigms other than the one examined here. These models are the configural-cue adaptive network model (Gluck & Bower, 1988b), a two-layer backpropagation network model, and ALCOVE (attention learning covering map, an exemplar model instantiated within a connectionist architecture; see Kruschke, 1992). In the simulations that follow, we first show that none of the models in unchanged form fit the human data on rule-learning data well. Second, we show that ALCOVE can fit the rule-learning data almost perfectly (the configural-cue and two-layer backpropagation models can fit moderately) when some of the prior biases specified in Bourne's (1974) inference model are translated in a principled way into a particular weight structure (and within which the particular links modified during learning are constrained in the two-layer backpropagation model). Finally, we show that allowing the models to acquire a weight structure appropriate for a conjunctive rule before performing the rule-learning task (to mimic the observation that humans appear to have a conjunctive bias upon entering the experimental setting), without utilizing the inference model, does not provide an adequate fit of the rule-learning data. Before reporting the simulations, we describe the data set on which the models were evaluated. These data represent human learning performance for rules that were used extensively in the well-defined concepts that dominated concept-learning research from the late 1950s to the early 1970s.

Bourne's Experiments on Logical Rule Learning

Bourne and his colleagues conducted a long series of experiments to systematically investigate all of the "basic" logical rules (see Bourne, 1974, for a summary). This database, which contains not only overall learning rates, but also the detailed pattern of errors, provides a crucial benchmark for any theory of conceptual rule learning. Not only are the results stable (as indicated below) and detailed, but the basic procedures of the rule-learning task are compatible with those embodied in current classification-learning tasks. Moreover, the rules seem representative of the abstract relational information incorporated in many real-world concepts. As just one example, the inclusive disjunctive (either *x* or *y* or both must be present) helps express the concept of a strike in a baseball game (the pitch must be swung at, or it must cross the plate between shoulders and knees, or both). Finally, the processes underlying learning of well-defined concepts in the standard laboratory paradigm seem to reflect the processes operating when such concepts are embedded in richer, more natural contexts (e.g., the stock market; Kozminsky, Kintsch, & Bourne, 1981).

In Bourne's typical rule-learning experiment, each stimulus represented one of three possible values (attributes) for each of four dimensions (color, shape, size, and number). Two defining attributes (e.g., large, triangle) were identified for the subjects at the outset, and the subjects were instructed to classify the stimuli into two categories, positive and negative, according to a rule relating the attributes to the concept. On a trial, a stimulus was presented, and a subject, given as much time as was needed, categorized it as a positive or negative example. The subject then received immediate feedback regarding the correctness of the response, and the next stimulus was presented after a 5-sec intertrial interval.

Using this procedure, Salatas and Bourne (1974) investigated learning for eight rules: four primary rules and four complementary rules. The primary rules were conjunctive (AND), inclusive disjunctive (OR), conditional (IF), and biconditional (IF AND ONLY IF). The four complementary rules were obtained by negating the primary rules (e.g., the complementary rule of the inclusive disjunctive is that *x* and *y* must both be absent), and were labeled alternative denial, joint denial, exclusive, and the exclusive disjunctive, respectively. The exclusive disjunctive rule is often called the exclusive-or rule. Table 1 shows correct responses to different types of stimuli for each rule, with "+" standing for the positive category, and "−" standing for the negative category. Types of stimuli are defined by the presence or absence of the critical attributes: TT = both attributes present, TF = Attribute 1 present and Attribute 2 absent, FT = Attribute 1 absent and Attribute 2 present, and FF = neither attribute present.

Table 1 also contains the mean number of errors before the learning criterion was reached (12 successive correct responses) for each rule (Salatas & Bourne, 1974). The distribution of errors across four different types of

Table 1
Errors for the Eight Bidimensional Rules by Human Subjects

Rules	Stimulus Types TT	TF	FT	FF	Mean Errors
Primary rules					
Conjunctive	0.33 (+)	1.08 (−)	0.92 (−)	0.33 (−)	2.67
Disjunctive	0.42 (+)	1.25 (+)	0.67 (+)	1.00 (−)	3.33
Conditional	1.25 (+)	7.50 (−)	2.42 (+)	8.92 (+)	20.08
Biconditional	1.75 (+)	8.08 (−)	6.08 (−)	11.33 (+)	27.25
Complementary rules					
Alternative denial	5.58 (−)	2.83 (+)	2.67 (+)	2.67 (+)	13.75
Joint denial	4.92 (−)	6.50 (−)	3.50 (−)	6.50 (+)	21.42
Exclusive	5.17 (−)	3.25 (+)	2.58 (−)	1.67 (−)	12.67
Exclusive disjunctive	5.33 (−)	4.25 (+)	5.17 (+)	3.42 (−)	18.17

Note—The data are adapted from Salatas and Bourne (1974). Plus and minus signs represent the correct categories, positive and negative, for each rule.

stimuli is also shown. The major finding to focus on at this point is that there are clear differences in the ease with which the rules were learned (with the conjunctive the easiest, and the biconditional the most difficult). These differences have proven stable: The ordering for the primary rules is well established (Bruner, Goodnow, & Austin, 1956; Haygood & Bourne, 1965; Hunt & Hovland, 1960; Neisser & Weene, 1962), and the ordering for all eight rules was essentially replicated by Neuman (1973).

Two additional points are noteworthy. First, a similar ordering is obtained even in the more general concept-learning paradigm in which subjects know neither the relevant features nor the rule (Neisser & Weene, 1962). Second, it has not proved possible to explain in a straightforward manner the factors that underlie the ordering of rule difficulty. Formulations such as those based on the number of primitive operations embedded in the rule (Neisser & Weene, 1972), a positive category focus strategy (see Bourne, 1974), and the informativeness of positive (or negative) instances (see Bourne, Dominowski, & Loftus, 1979) have not adequately captured the data. Thus, these concept-learning data pose a nontrivial challenge for models of concept acquisition.

General Simulation Procedure

In performing the simulations, we attempted to follow Salatas and Bourne's (1974) experimental procedure closely. Prominent in our considerations were the number of trials and the stimulus sequence, which were kept the same for all the simulations reported. We used 160 trials because Salatas and Bourne used 160 for the maximum number of trials in learning a rule. The complete presentation sequence for the stimuli was obtained by counterbalancing the order of presentation of the four input types of stimuli (TT, TF, FT, and FF) and counterbalancing the stimuli within each of the TF, FT, and TF input types, which yielded a sequence of 64 stimuli; this sequence was then repeated until the number of trials reached 160. Salatas and Bourne used a 40-stimulus sequence for each problem (repeating it if necessary), with the constraints that it contain 10 stimuli from each stimulus type, that the first 4 stimuli include 1 stimulus from each stimulus type, and that the 1st stimulus be a TT stimulus. Those constraints were observed in the first 40 trials of our sequence.

The procedure for testing the models (the configural-cue models, two-layer backpropagation models, and ALCOVE models) was basically the same. With initial values of parameters set by the modeler, the models performed for 160 trials for each rule. On each trial, each model was presented with a stimulus, then produced a classification choice probability, and received feedback about the correct category. The choice probabilities were derived from the activation of output nodes. Using these choice probabilities, we calculated the probabilities of reaching the learning criterion at Trials 12–160 (the stopping criterion was 12 consecutive correct responses in Salatas & Bourne, 1974). These stopping probabilities were used in concert with the choice probabilities to compute the expected number of errors for each type of stimulus for the eight rules. Then these 32 predicted data points (8 rules × 4 stimulus types) were compared with 32 data points from human subjects (see the Appendix for details). A nonlinear optimization algorithm was used to find a set of parameters that minimized the squared difference between the predicted data and the humans' data.

Our first attempt to test these models was to have the models learn the eight rules and compare the error distributions produced by the models with those found for humans. For the two-layer backpropagation model, small random weights (ranging from −0.25 to +0.25) were assigned to all the links as initial weights. For the configural-cue model and the ALCOVE model, zero weights were assigned (further details of the models can be found in the following sections). The results were not encouraging. Table 2 shows that none of the models were able to reproduce the ordering of rule difficulty displayed by human subjects.[1] Moreover, the configural-cue model, the two-layer backpropagation model, and the ALCOVE model accounted for only 14.2%, 18.2%, and 68.6% of the variance in the human data, respectively. Notice that the configural-cue model and ALCOVE predicted the same results for the primary and complementary rules because a particular complementary rule partitions the exemplars into the same subgroupings as does its respective primary rule. The only difference between the two is that the labels applied to each subgrouping are reversed. This difference, however, is transparent to the configural-cue model and ALCOVE.

There are at least two explanations for why the formal "off-the-shelf" concept-learning models considered here failed to account for the learning patterns of subjects faced with rule-based concepts. As mentioned earlier, these associationist models may not capture the processes involved in rule-based learning tasks. An alternative is that such models are adequate in principle, but require modification to reflect essential contributions of prior knowledge. A number of researchers have noted that subjects appear to adopt a conjunctive set at the outset of rule-oriented concept-learning tasks (e.g., Bruner et al., 1956; Medin, Wattenmaker, & Michalski, 1987), and Bourne (1974) explicitly linked the pattern of rule difficulty to this conjunctive set. It is possible that current models could provide an accurate representation of the data if appropriate "prior knowledge" or biases were incorporated into the models. In the remainder of this paper, we will explore this possibility.

In incorporating prior biases, we relied mainly on Bourne's (1974) inference model of conceptual rule learning. Bourne's inference model states that human subjects come to the experiment with four prior biases. First, TT stimuli are assigned to the positive category (Bias 1). Second, FF stimuli are assigned to the negative category (Bias 2). Third, TF and FT stimuli are assigned to the category to which FF stimuli are assigned (Bias 3). Finally, TT and FF stimuli are assigned to different categories (Bias 4). In the following simulations, we try to incorporate these biases into the configural-cue, two-layer backpropagation, and ALCOVE models.

Table 2
Actual Human Errors and Predicted Errors Produced by Unmodified Models

Rules	Human Data	Config	Backpr	ALCOVE
Primary rules				
Conjunctive	2.67	12.12	2.56	8.77
Disjunctive	3.33	15.69	4.52	14.47
Conditional	20.08	11.62	9.64	14.68
Biconditional	27.25	19.90	15.13	21.96
Complementary rules				
Alternative denial	13.75	12.12	10.72	8.77
Joint denial	21.42	15.69	4.72	14.47
Exclusive	12.67	11.62	5.09	14.68
Exclusive disjunctive	18.17	19.90	18.18	21.96
R^2		.142	.182	.686

Note—The scores of R^2 are based on all 32 data points (see Table 3; the R^2 is the proportion of variance accounted for by the model). Config = configural-cue model; Backpr = two-layer backpropagation model. The human data are adapted from Salatas and Bourne (1974).

Configural-Cue Model

The configural-cue model is a one-layer linear network model, with some of the input nodes representing configural cues (Gluck & Bower, 1988b; Gluck, Bower, & Hee, 1989). In the present stimulus domain, there were nine configural input nodes that represented the unique combination of the two critical stimulus attributes (e.g., [large, medium, small] × [triangle, square, hexagon]). In addition, there were six input nodes for the individual attributes of stimuli. Thus the model had 15 input nodes connected to two output nodes.

The first three input nodes coded for the dimension of size.[2] The first node represented the relevant size, and it was activated when the relevant size was present in the stimulus. The second and third input nodes represented irrelevant sizes; they were activated when a corresponding irrelevant size was present in the stimulus. The fourth, fifth, and sixth nodes coded for the dimension of shape. The fourth node represented the relevant shape, and the fifth and sixth nodes represented irrelevant shapes. Their activation behaved in a fashion paralleling that described for the size input nodes.

The 7th–15th nodes represented configural cues. One of these input nodes represented stimuli that had both the relevant size and the relevant shape (TT stimulus type). This TT node was turned on when a TT stimulus was presented, and it was turned off otherwise. Another two input nodes represented stimuli that had the relevant size but an irrelevant shape (TF stimulus type). The next two input nodes represented stimuli that had an irrelevant size but the relevant shape (FT stimulus type). The last four of these configural input nodes represented stimuli that had both an irrelevant size and an irrelevant shape (FF stimulus type). The two output nodes represented the positive category and the negative category, respectively.

To reflect subjects' tendencies as they begin the experiment (cf. Bourne, 1974), we preset the model toward a conjunctive rule, and the weights were used as the initial weights for the target problems. Because a conjunctive tendency could be instantiated in several different ways, only one of which might capture the bias presumably displayed by humans, we attempted to instantiate Bourne's specification of subjects' microbiases mediating the conjuncture set in the architecture of the configural-cue model. This biasing was achieved by making the node representing the TT configural cue excite the positive-category output node and inhibit the negative-category output node (Bias 1), and by making the nodes representing the TF, FT, and FF configural cues inhibit the positive-category output node and excite the negative-category output node (Bias 2). The nodes representing individual attributes were set equal to zero.[3] Biases 3 and 4 were not implemented because the configural-cue model, at least in the current one-layer architecture, could not capture the interacting relations among different types of stimuli. Eleven parameters were used: nine for setting the initial weights from the input nodes to the output nodes, one for the learning rate, and one (ϕ in Equation 2) for mapping the activations of the output nodes onto the response probabilities.

In this one-layer network, activation on output node k is given by

$$a_k^{\text{out}} = \sum_j w_{jk} a_j^{\text{in}}, \tag{1}$$

where w_{jk} is the weight from input node j to output node k, and a_j^{in} is the activation of input node j. The probability that the kth category response is made is given by

$$P_k = \exp(\phi a_k^{\text{out}}) / \sum_k \exp(\phi a_k^{\text{out}}), \tag{2}$$

where ϕ is a mapping constant. The learning rule used to update weights is given by

$$\Delta w_{jk} = \lambda a_j^{\text{in}} (t_k - \sum_j w_{jk} a_j^{\text{in}}), \tag{3}$$

where Δw_{jk} is weight change, λ is a learning rate, and t_k is a teaching signal, which was set to +1 for the correct category and −1 for the incorrect category.

The model was presented with the stimulus set, and it produced output activations on the output nodes following each stimulus presentation. The stimulus presentation sequence comprising the 160 trials was that described in the General Simulation Procedure. We estimated learning errors for the model as follows. For each trial, a choice probability was computed according to Equation 2. Then, identically to the procedure described in the General Simulation Procedure, the choice probabilities for all 160 trials were used to compute the expected number of errors. We used a nonlinear optimization algorithm to find a set of parameters that minimized the squared difference between the predicted and the humans' data, and that satisfied the biases specified by the modeler.

Results. Overall, the configural-cue model provided a reasonably good fit accounting for 86.8% of the variance in the 32 data points obtained from human subjects and correctly producing the relative difficulty within the set of primary rules and within the set of complementary rules (see Table 3). However, the model made fewer errors in learning the conditional than in learning the exclusive disjunctive, which is the reverse of the humans' data. With regard to the relative difficulty of stimulus types within rules, the configural-cue model again provided a generally good fit, except for the alternative denial and exclusive disjunctive (see Table 4). The correlations between the predicted and observed errors across the four stimulus types were .71, .71, 1.00, .99, .21, .86, .68, and −.04 for the conjunctive, inclusive disjunctive, conditional, biconditional, alternative denial, joint denial, exclusive, and exclusive disjunctive, respectively. With regard to the relative difficulty of a stimulus type collapsed across rules, the model provided an excellent fit. The correlations between the predicted and observed errors across the eight rules were .97, .96, .93, and .95 for the TT, TF, FT, and FF stimulus types, respectively. In sum, building prior conjunctive bias into the configural-cue model substantially improved both the variance accounted for and the accuracy of the rule-difficulty ordering. Nonetheless, the model was not perfectly successful; it did not completely produce the qualitative patterns observed for human learners—that is, the order of rule difficulty.

Configural-Cue Model With Conjunctive Training

One might argue that Bourne's inference model was not necessary to simulate Salatas and Bourne's (1974) data, and that a more successful simulation could be obtained by just letting the network learn the initial tendency favoring the conjunctive rule, and then treating this tendency as the bias that subjects bring to the experiment. To test this idea, we trained the configural-cue model to produce the same initial output activations (at the outset of learn-

Table 3
Actual Human Errors and Errors Produced by All Three Models With Bias

	Human	Modeler-Constructed Bias			Trained Bias		
Rules	Data	Config	Backpr	ALCOVE	Config	Backpr	ALCOVE
Primary rules							
Conjunctive	2.67	2.89	4.31	0.71	2.88	1.32	4.26
Disjunctive	3.33	4.38	5.18	3.63	4.38	2.35	5.23
Conditional	20.08	17.98	18.85	19.51	17.98	38.63	13.90
Biconditional	27.25	27.11	27.73	27.37	27.11	9.85	24.54
Complementary rules							
Alternative denial	13.75	13.98	13.20	14.64	13.98	17.38	17.14
Joint denial	21.42	23.21	21.38	20.91	23.21	18.03	24.36
Exclusive	12.67	11.00	13.67	12.55	10.99	44.47	12.04
Exclusive disjunctive	18.17	18.46	17.10	18.14	18.45	5.05	19.36
R^2		.868	.841	.987	.868	.224	.943
No. of parameters		11	10	8	2*	3*	4*

Note—The scores of R^2 are based on all 32 data points (see Table 4; the R^2 is the proportion of variance accounted for by the model). Config = configural-cue model; Backpr = two-layer backpropagation model. The human data are adapted from Salatas and Bourne (1974). *These numbers of parameters are not directly comparable with those for models with modeler-constructed bias, because training plays a role similar to that which parameters (9, 8, and 4 for the configural-cue, two-layer backpropagation, and ALCOVE models, respectively) used for the modeler-constructed bias do.

Table 4
Error Distribution Produced by All Three Models With Modeler-Constructed Bias

	Stimulus Types											
	Configural-Cue				Two-layer Backpropagation				ALCOVE			
Rules	TT	TF	FT	FF	TT	TF	FT	FF	TT	TF	FT	FF
Primary rules												
Conjunctive	0.20	1.28	0.64	0.77	0.47	2.39	1.22	0.23	0.12	0.41	0.15	0.03
Disjunctive	0.03	2.26	1.53	0.57	0.51	1.65	2.40	0.62	0.08	1.16	1.42	0.97
Conditional	0.95	6.78	1.44	8.81	1.89	7.36	4.06	5.54	0.14	7.88	2.10	9.37
Biconditional	2.26	8.57	5.19	11.09	3.11	6.58	5.56	12.48	1.74	8.06	6.52	11.05
Complementary rules												
Alternative denial	3.93	2.43	2.57	5.05	4.47	2.07	2.50	4.16	4.98	2.62	3.44	3.60
Joint denial	3.31	6.99	3.43	9.47	5.27	5.01	3.53	7.56	5.52	5.14	4.30	5.95
Exclusive	3.72	3.75	1.29	2.25	5.86	4.01	3.08	0.71	4.49	2.46	2.44	3.16
Exclusive disjunctive	4.36	5.81	4.27	4.01	6.49	3.06	4.57	2.98	5.66	3.10	4.62	4.77

Note—Refer to Table 1 for comparable human data.

ing) as those produced by the configural-cue model with prior biases, and then used this state as the initial starting bias.

This conjunctive-trained network had the same number of input nodes and output nodes as did the configural-cue model with prior bias that we tested. To reflect the initial conjunctive bias that subjects apparently bring to the experiment, initial weights were obtained by training the network until it produced the same output activations on output nodes (for nine input stimuli, respectively) as did the model with prior bias at the outset. The conjunctive training played the same role as did the eight parameters of the previous model that were used to set the prior biases. In order to optimize the chance of finding the best fit that this conjunctive-trained model would support, two extra parameters were used: one for the learning rate, and one for the mapping constant.

Results. The conjunctive-trained configural-cue model did not improve the fit: It accounted for 86.8% of the variance in the 32 data points for human subjects, as did the configural-cue model with prior bias. The qualitative fit in terms of predicting rule difficulty was unchanged as well (see Table 3). In sum, regardless of whether the bias was acquired by the model or set by the theorist, the model still failed to completely produce the order of rule difficulty manifested by subjects.

Two-Layer Backpropagation Model With Prior Bias and Structural Constraints

Some theorists have argued that a one-layer network model is too simplistic to support complex learning (Minsky & Papert, 1969; Rumelhart & McClelland, 1986). Therefore, we next examined two-layer network models with nine input nodes, various numbers of hidden nodes, and two output nodes (9:*n*:2 models). The nine input nodes represented nine input stimuli (e.g., [large, medium, small] × [triangle, square, hexagon]), respectively. The two output nodes represented the positive and the negative categories. We chose to report the following model, which has three hidden nodes, because it was the most successful of the backpropagation models.

As described earlier, Bourne's (1974) inference model specifies four initial biases that presumably mediate subjects' initial tendencies to adopt a conjunctive set. Borrowing from Bourne's ideas, we structured a network at the outset to reflect the particular collection of tendencies outlined in Bourne's inference model. This network model is diagrammed in Figure 1. Bias 1 of Bourne's inference model (for a description of the biases, see the General Simulation Procedure above) was implemented by having the TT input node exclusively excite the first hidden node (representing the TT stimuli), which excited the positive output node and inhibited the negative output node. Bias 2 was implemented by having the FF input nodes excite the third hidden node (representing the FF stimuli), which excited the negative output node and inhibited the positive output node. Bias 3 was implemented by having the TF and FT input nodes excite the third hidden node. Bias 4 was implemented by having the FF in-

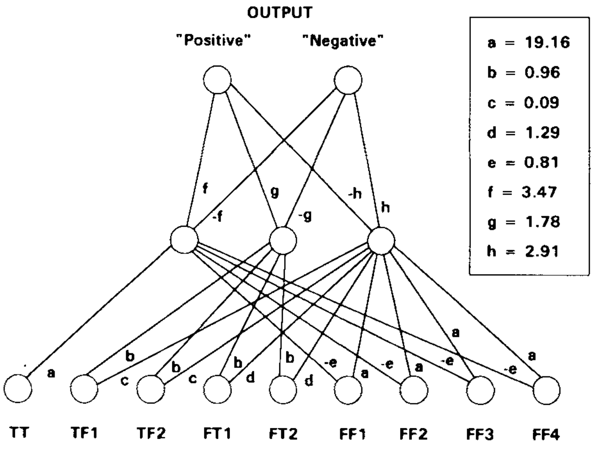

Figure 1. A two-layer network model with prior bias and structural constraints. (The numbers indicate best-fitting initial weights.)

put nodes inhibit the first hidden node. Given the short period of time during which the subjects experience the experimental materials, we assume that the organizational tendencies do not change; that is, no learning occurs in the input-hidden links. Ten parameters were used: eight for setting the weights, one for the learning rate of the hidden-output layer, and one for the mapping constant (ϕ in Equation 5).

In this two-layer network, activation on output node k is given by

$$a_k^{\text{out}} = \textstyle\sum_j w_{jk} a_j^{\text{hid}}, \tag{4}$$

where w_{jk} is the weight of the link from hidden node j to output node k, and a_j^{hid} is the activation of hidden node j. The probability that the kth category response is made is given by

$$P_k = 1/[1+\exp(-\phi a_k^{\text{out}})]. \tag{5}$$

The learning rule was the delta rule modified to include a momentum term (Rumelhart et al., 1986), which was used to increase learning rate without oscillation so that

$$\Delta w_{jk}(t) = \lambda a_j^{\text{hid}}(t_k - a_k^{\text{out}})a_k^{\text{out}}(1-a_k^{\text{out}}) + \beta \Delta w_{jk}(t-1), \tag{6}$$

where $\Delta w_{jk}(t)$ is weight change at trial t, λ is a learning rate, t_k is a teaching signal, which was set to $+1$ for the correct category and 0 for the incorrect category, and β is a momentum term that was set to 0.9. The setting of β to 0.9 is typical in simulations by Rumelhart et al. (1986).

Activation on hidden node j has a range of -1 to $+1$ and is given by

$$a_j^{\text{hid}} = 2/[1+\exp(-\textstyle\sum_i w_{ij} a_i^{\text{in}})] - 1, \tag{7}$$

where w_{ij} is the weight of the link from input node i to hidden node j, and a_i^{in} is the activation on input node i.

Results. The two-layer backpropagation model with both a conjunctive bias and structural constraints provided a reasonably good fit to the Salatas and Bourne (1974) human rule-learning data (compare columns 1 and 4 of Table 3). The model accounted for 84.1% of the variance in the 32 data points obtained from human subjects, and it correctly ordered the primary rules in terms of difficulty. Only the alternative denial and the exclusive were reversed by the model.

The structurally constrained network model also provided a reasonably good fit for individual stimulus types (see Table 4). With regard to the relative difficulty of stimulus types within rules, the network model provided a good fit except for the inclusive disjunctive rule. The correlations between the predicted and observed errors across the four stimulus types were .93, .18, .85, .94, .63, .77, .96, and .85 for the conjunctive, inclusive disjunctive, conditional, biconditional, alternative denial, joint denial, exclusive, and exclusive disjunctive, respectively. With regard to the relative difficulty of a stimulus type collapsed across rules, the structurally constrained model provided an excellent fit. The correlations between the predicted and observed errors across the eight rules were .95, .93, .90, and .93 for the TT, TF, FT, and FF stimulus types, respectively.

Two-Layer Backpropagation Model With Conjunctive Training

To test the idea that conjunctive training without utilizing Bourne's inference model would be enough to simulate the human concept-learning data, we trained the two-layer backpropagation model to produce the same initial output activations as those produced by the two-layer backpropagation model with prior bias, and then used this state as the initial bias. The conjunctive-trained backpropagation model had the same number of input nodes, hidden nodes, and output nodes as did the backpropagation model with prior biases and structural constraints.

Unlike the backpropagation model with prior bias, for the conjunctive-trained model learning occurred in the first layer. The learning rule for updating weights from input nodes to hidden nodes is given by

$$\Delta w_{ij}(t) = 2\lambda a_i^{\text{in}} a_j^{\text{hid}}(1-a_j^{\text{hid}})\textstyle\sum_k[(t_k - a_k^{\text{out}})a_k^{\text{out}}(1-a_k^{\text{out}})w_{jk}] + \beta\Delta w_{ij}(t-1). \tag{8}$$

In addition to the conjunctive training that replaced the eight parameters of the two-layer backpropagation model with prior bias to set prior biases, three extra parameters were used: two for the learning rates, one for the mapping constant.

Results and Discussion. The conjunctive-trained backpropagation model showed a dramatic decrease in the fit of the human concept-learning data (see Table 3). The exclusive, which was the third easiest for humans, was the most difficult for the model. The biconditional, which was the most difficult for humans, was the fourth easiest for the model. The model accounted for only 22.4% of the variance observed in the humans' data. Thus, simple conjunctive training of the hidden-node network was not sufficient to simulate the humans' data. This may not be surprising, given that human learners have almost certainly not acquired their presumed conjunctive bias on the basis of experience with one laboratory concept problem. Perhaps if one could specify the experiences (training) by which humans acquire the biases that enter into the rule-learning task, a hidden-node network could be trained to capture the knowledge state at which humans enter the rule-learning task.

The fact remains that a faithful implementation of Bourne's (1974) inference model into a network with hidden nodes or into the existing configural-cue model can produce a reasonably good fit of the humans' rule-learning data. Still, neither model perfectly captured the ordering of rule difficulty evidenced by subjects.

Next, we explore ALCOVE (Kruschke, 1992), which combines the error-driven learning of networks with the successful exemplar-oriented representations of previous categorization models (see, e.g., Medin & Schaffer, 1978; Nosofsky, 1986). With these features, ALCOVE can account for many effects of categorization: It can learn to attend to relevant stimulus dimensions and to correlated stimulus dimensions, it is not prone to catastrophic interference, and it shows three-stage learning of rules and exceptions (Kruschke, 1992).

ALCOVE With Prior Bias

In the current stimulus domain, there were two input nodes (stimulus dimension), nine hidden (exemplar) nodes, and two output (category) nodes. The first node represented the dimension of size; the second input node represented the dimension of shape. The three values of each dimension were coded as −1, 0, and +1. Relevant attributes were coded as 0, and irrelevant attributes were coded as −1 and +1.[4] The first two biases of Bourne's (1974) inference model were implemented in the hidden-output layer: The TT exemplar excites the positive output node, and inhibits the negative output node (Bias 1). The four FF exemplars inhibit the positive output node and excite the negative output node (Bias 2).

Biases 3 and 4 were not implemented, because the implementation was not possible with ALCOVE's activation function of exemplar nodes (Equation 9) that was an important element of ALCOVE. Eight parameters were used: two for the learning rates of attention strengths and association weights, one for the specificity constant (c in Equation 9), one for the mapping constant (ϕ in Equation 11), and four for setting the initial weights from hidden nodes to output nodes.

For a given stimulus, the activation of the jth hidden node is given by

$$a_j^{\text{hid}} = \exp[-c(\textstyle\sum_i \alpha_i |h_{ji} - a_i^{\text{in}}|^r)^{q/r}], \tag{9}$$

where c is the specificity constant of the node, α_i is an attention strength for the ith dimension, h_{ji} is the value of the ith dimension for the hidden node j, and a_i^{in} is the value of the ith dimension for the input stimulus. In this simulation, we used a city-block metric ($r = 1$) with exponential similarity gradient ($q = 1$).

The activation of output node k is given by

$$a_k^{\text{out}} = \textstyle\sum_j w_{jk} a_j^{\text{hid}}, \tag{10}$$

where w_{jk} is the association weight of the link from hidden node j to output node k. The output activations are mapped onto the response probabilities by

$$P_k = \exp(\phi a_k^{\text{out}}) / \textstyle\sum_k \exp(\phi a_k^{\text{out}}), \tag{11}$$

where P_k is the probability that the kth category is chosen, and ϕ is a mapping constant.

As learning occurs, the association weights and attention strengths are, respectively, updated by

$$\Delta w_{jk} = \lambda_w (t_k - a_k^{\text{out}}) a_j^{\text{hid}}, \tag{12}$$

where λ_w is a learning rate for the association weights, and

$$\Delta \alpha_i = -\lambda_\alpha \textstyle\sum_j [\sum_k (t_k - a_k^{\text{out}}) w_{jk}] a_j^{\text{hid}} c |h_{ji} - a_i^{\text{in}}|, \tag{13}$$

where λ_α is a learning rate for the attention strengths, and t_k is a teaching signal at output node k. The teaching signal is set to the maximum of +1 and a_k^{out} for a correct category, and to the minimum of −1 and a_k^{out} for an incorrect category.

On each trial, the model was presented with a stimulus, and it produced a choice probability. The choice probabilities for all 160 trials were then used to compute the expected number of errors as described in the General Simulation Procedure.

Results. ALCOVE with prior bias perfectly simulated the relative difficulty of both the primary and the complementary rules (see Table 3). ALCOVE accounted for 98.7% of the variance in the 32 data points obtained from human subjects. With regard to the relative difficulty of stimulus types within rules, the model provided a reasonably good fit (see Table 4). The correlations between the predicted and observed errors across the four stimulus types were .82, .59, 1.00, 1.00, .87, .76, .71, and .39 for the conjunctive, inclusive disjunctive, conditional, biconditional, alternative denial, joint denial, exclusive, and exclusive disjunctive, respectively. With regard to the relative difficulty of a stimulus type collapsed across rules, ALCOVE provided an excellent fit. The correlations between the predicted and observed errors across the eight rules were .98, .98, .95, and .98 for the TT, TF, FT, and FF stimulus types, respectively.

ALCOVE With Conjunctive Training

To test the idea that conjunctive training without consideration of Bourne's (1974) inference model would be enough to simulate the human concept-learning data, we trained ALCOVE to produce the same initial output activations as those produced by ALCOVE with prior bias, and then used this state as the initial bias. The conjunctive-trained ALCOVE model had the same number of input nodes, hidden nodes, and output nodes, and the same value of the specificity constant to produce the same initial bias. In addition to the conjunctive training that replaced the four parameters used to set biases for ALCOVE with prior bias, three extra parameters were used: two for the learning rates, and one for the mapping constant.

Results. The conjunctive-trained ALCOVE accounted for 94.3% of the variance in the human data. Although the fit in terms of accounted-for variance decreased by only 4.4% relative to the version with prior bias, the qualitative fit to the order of rule difficulty declined. Examination of Table 3 shows that the conditional and the biconditional now become indistinguishable in difficulty and that the exclusive and the exclusive disjunctive also become indistinguishable, whereas for humans, the conditional is easier than the biconditional, and the exclusive is easier than the exclusive disjunctive.

General Discussion

We examined a configural-cue model, a two-layer backpropagation model, and an exemplar-based model (ALCOVE) in terms of their ability to mimic human learning of well-defined concepts based on logical rules. We tried to test these models under conditions that closely approximated the experimental paradigm under which the human data were originally collected, so that the success or failure of the models could not be assailed in terms of a poor match between the learning conditions for the model and those for the human learners. The results were quite encouraging regarding the ability of at least one model—ALCOVE—to incorporate biases that human learners presumably bring to the rule-learning situation, given sufficient input into the network's weights from the modeler. This finding significantly extends the range of category-learning effects that can be accounted for by ALCOVE. The empirical results to which ALCOVE has been applied (and for which models of this type have been targeted) derive from either (1) paradigms in which the primary task is to ascertain the relevant stimulus dimensions used for classification or (2) so-called fuzzy categories, in which there may be no easily explicated rule that defines the category boundaries (see Kruschke, 1992). The data considered here were from a paradigm in which dimensional learning was purposefully circumvented (by providing learners with the relevant attributes) to reveal the learning process involved in abstracting the rule that governs classification (Bourne, 1967). The success of ALCOVE in capturing the data in the rule-learning paradigm adds to its appeal as a possible foundation for a comprehensive category-learning model.

It is instructive to compare ALCOVE with the other models, to get possible insight into why ALCOVE fits better than the other models. One distinction is that ALCOVE can adjust attention strengths that are used to discriminate among different stimuli. (A larger attention strength for a dimension means that the model pays more attention to the dimension. A larger average of attention strengths of the involved dimensions produces more discriminable stimuli for the model.) We found that when ALCOVE, with prior bias, learns the complementaryrules (alternative denial, joint denial, exclusive, and exclusive disjunctive), attention strengths become about three or four times larger than they do when it learns the primary rules (conjunctive, inclusive disjunctive, conditional, and biconditional). In other words, to learn the complementary rules, ALCOVE individuates among stimuli much more than it does when it learns the primary rules. On the other hand, the two-layer backpropagation and configural-cue models, lacking a mechanism to change the amount of attention during learning, cannot change the stimulus discriminability. It might be the capability to adapt attention (i.e., stimulus discriminability) that gives ALCOVE an advantage in fitting the rule-learning data.

Despite this apparent advantage, allowing ALCOVE to acquire its own conjunctive bias did not appear to capture the nature of the bias that human subjects apparently come into the rule-learning experiments with (the same held for the configural-cue and two-layer backpropagation models). Thus, there appear to be several different ways in which one might represent a conjunctive bias, not all of which reflect what human learners do (see Pavel, Gluck, & Henkle, 1988, for a similar point). More generally, merely making an adaptive network mimic superficial response tendencies observed in experimental data (e.g., in the present work, the initial conjunctive tendencies) may not necessarily constrain the network enough to allow it to acquire representations and processes underlying human concept learning. This point reinforces McCloskey's (1991) assertion that to allow an adaptive network to build itself is less preferable as a modeling technique than to use such networks as a tool to formalize the theorist's explicit assumptions and intuitions about the structures underlying the process of interest.

Interestingly, however, implementation of the successful ALCOVE with bias did not require explicit inclusion of all the prior biases hypothesized by Bourne (1974). Bourne's inference model assumes that subjects have biases such that TF and FT stimuli belong to the category where FF stimuli are placed (Bias 3) and that TT and FF belong to opposing categories (Bias 4). In ALCOVE, Bias 3 is operating only when the perceived similarity among the TF, FT, and FF stimuli is relatively high (i.e., low attention strengths). ALCOVE does not have Bias 4 at all. It seems, then, that Bourne's theoretical account and the ALCOVE-with-bias model might produce different predictions regarding aspects of behavior not captured in the mean error data (i.e., behaviors relating to the relations specified in Biases 3 and 4). If so, ALCOVE with bias could provide a modified theory for Bourne's inference model of conceptual rule learning. The ALCOVE model could also be used to detail the transitions from the learner's initial tendencies to the eventual structure that mediates criterial performance, something that has not been achieved so far.

More generally, the present work demonstrates that an incremental, associative learning model can account for learning on concept problems for which hypothesis-testing models have been traditionally favored. Indeed, such models are still favored, as indicated by a recent computational model of rule-based concepts that features hypothesis testing (Pazzani, 1991). We may need to reexamine assumptions regarding subjects' use of hypothesis-testing procedures in learning well-defined rules (see also Kellogg & Bourne, 1989), because ALCOVE was successful in recapitulating rule-learning data despite its lack of mechanism for hypothesizing and testing rules (Kruschke, 1992, p. 40). On the other hand, we suggest that the incremental learning formalized in adaptive networks is not necessarily incompatible with hypothesis-testing processes. One might view our zero-weighted associations in the backpropagation model between input and hidden nodes as a hypothesis regarding which associations are most relevant for the task presented to the model. The suggestion here is that the particular network configuration that we have described might be viewed as a formal approximation of the hypotheses or inferential tendencies that subjects form after receiving the particular instructions associated with the experimental task. If significant aspects of the instructions and stimuli were to change (as, e.g., in Pazzani's, 1991, paradigm), the configuration of the particular network would necessarily change to reflect changes in the hypotheses. An example of how one might formalize the mechanism(s) that coordinate the formulation and dynamic selection of such networks (hypotheses) can be found in Busemeyer and Myung (1992).

REFERENCES

Bourne, L. E., Jr. (1967). Learning and utilization of conceptual rules. In B. Kleinmuntz (Ed.), *Concepts and the structure of memory* (pp. 1-32). New York: Wiley.

Bourne, L. E., Jr. (1974). An inference model of conceptual rule learning. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 231-256). Potomac, MD: Erlbaum.

Bourne, L. E., Jr., Dominowski, R. L., & Loftus, E. F. (1979). *Cognitive processes*. Englewood Cliffs, NJ: Prentice-Hall.

Bourne, L. E., Jr., & Guy, D. E. (1968). Learning conceptual rules: I. Some interrule transfer effects. *Journal of Experimental Psychology*, **76**, 423-429.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

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

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, **121**, 177-194.

Estes, W. K., Campbell, J. A., Hatsopoulus, 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.

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. R. (1989). A configural-cue network model of animal and human associative learning. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Haygood, R. C., & Bourne, L. E., Jr. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review*, **72**, 175-195.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

Homa, D. (1984). On the nature of categories. In G. B. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 49-94). New York: Academic Press.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 418-439.

Horton, D. L., & Turnage, T. W. (1976). *Human learning*. Englewood, NJ: Prentice-Hall.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs*, **28**(1, Whole No. 123).

Hunt, E. B., & Hovland, C. I. (1960). Order of consideration of different types of concepts. *Journal of Experimental Psychology*, **59**, 220-225.

Kellogg, T. T., & Bourne, L. E., Jr. (1989). Nonanalytic-automatic abstraction of concepts. In J. B. Sidowski (Ed.), *Conditioning, cognition, and methodology: Contemporary issues in experimental psychology* (pp. 89-111). Lanham, MD: University Press of America.

Kozminsky, E., Kintsch, W., & Bourne, L. E., Jr. (1981). Decision making with texts: Information analysis and schema acquisition. *Journal of Experimental Psychology: General*, **110**, 363-380.

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

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, **2**, 387-395.

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

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, **11**, 299-339.

Minsky, M. L., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

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

Neuman, P. G. (1973). *Directional and neutral category labels in bidirectional concept identification problems*. Unpublished master's thesis, University of Colorado, Boulder.

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

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

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, **19**, 131-150.

Pavel, M., Gluck, M. A., & Henkle, V. (1988). Constraints on adaptive networks for modeling human generalization. In *Proceedings of the November 1988 Neural Information Processing Systems Conference*. Hillsdale, NJ: Erlbaum.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 416-432.

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

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-233.

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

Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 110-146). Cambridge, MA: MIT Press.

Salatas, H., & Bourne, L. E., Jr. (1974). Learning conceptual rules: III. Processes contributing to rule difficulty. *Memory & Cognition*, **2**, 549-553.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 433-443.

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

NOTES

[1] We also varied the number of hidden nodes for the backpropagation model (9:3:2, 9:8:2), but could not get a satisfactory fit. The performances of these models are available on request.

[2] For purposes of exposition, the two relevant dimensions will be specified as size and shape. This is an arbitrary choice; the particular dimensions are irrelevant to the models.

[3] It may seem more natural to make the relevant-attribute nodes excite the positive-category output node and inhibit the negative-category output node, and to make the irrelevant-attribute nodes inhibit the positive-category output node and excite the negative-category output node. With this nonzero biasing, however, the configural-cue model produced a worse fit, accounting for 69.7% of variance in the 32 data points from human subjects.

[4] When relevant attributes were coded as -1 (or $+1$), and irrelevant attributes as 0 and $+1$ (or 0 and -1), results (not reported) were essentially the same, accounting for 97.3% of the variance of the humans' data. Although this coding seems reasonable for size and number dimensions, it may not work well for color and shape dimensions. (Are triangles and hexagons psychologically twice distant from each other as they are from squares?)

APPENDIX

Here, we will describe (1) how we derived the expected number of errors for each stimulus type and the expected number of total errors made during rule learning, and (2) how we estimated parameters for the models.

Define the predicted probability of a correct response on trial t as $P_{c,t}$. We obtain the probability of error on trial t from

$$P_{e,t} = 1 - P_{c,t}.$$

A subject always reaches Trial 12 because the criterion is 12 consecutive correct responses. The probability that the subject will stop at Trial 12 is given by

$$P(S_{12}) = \prod_{i=1}^{12} P_{c,i}.$$

Generally, the probability that the subject will stop at trial t (S_t), given that the subject has reached trial t (R_t), is given by

$$P(S_t \mid R_t) = \begin{cases} 0, & 1 \le t \le 11 \\ \prod_{i=t-11}^{t} P_{c,i}, & 12 \le t \le 159 \\ 1, & t = 160. \end{cases}$$

The probability that the subject will reach trial t is given by

$$P(R_t) = \begin{cases} 1, & 1 \le t \le 12. \\ [1 - P(S_{t-1} \mid R_{t-1})]P(R_{t-1}). & 13 \le t \le 160. \end{cases}$$

(Note that $P[R_{12}] = 1$.)

The probability that the subject will reach trial t and stop on trial t is given by

$$P(S_t \cap R_t) = P(R_t)P(S_t \mid R_t).$$

Now, the expected sum of errors contributed by Trial 1 is given by

$$E(e_1) = P_{e,1} \sum_{i=13}^{160} P(S_i \cap R_i).$$

Generally, the expected sum of errors contributed by trial t is given by

$$E(e_t) = P_{e,t} \sum_{i=t+12}^{160} P(S_i \cap R_i), \qquad 1 \le t \le 148.$$

The expected total of errors for each stimulus type (TT, TF, FT, and FF) of a rule is given by

$$E(e_{\text{stimulus type}}) = \sum E(e_i),$$

where the sum above extends across all trials, i, on which a particular stimulus type occurred.

The expected total of errors for a rule is given by

$$E(e_{\text{rule}}) = \sum E(e_{\text{stimulus type}}),$$

where the above sum extends across all four stimulus types.

Now we have 32 data points (4 stimulus types × 8 rules) predicted by a model, and 32 data points from human subjects. To estimate parameters, we obtain the sum of squared prediction errors by summing up the squared differences between the humans' data and the data predicted by the model. Then, using a nonlinear optimization algorithm, we find a set of parameters that minimizes the sum of squared prediction errors.

(Manuscript received July 29, 1992;
revision accepted for publication November 30, 1992.)