JOURNAL OF MEMORY AND LANGUAGE 27, 166–195 (1988)

Evaluating an Adaptive Network Model of Human Learning

MARK A. GLUCK AND GORDON H. BOWER

Stanford University

This paper explores the promise of simple adaptive networks as models of human learning. The least-mean-squares (LMS) learning rule of networks corresponds to the Rescorla–Wagner model of Pavlovian conditioning, suggesting interesting parallels in human and animal learning. We review three experiments in which subjects learned to classify patients according to symptoms which had differing correlations with two diseases. The LMS network model predicted the results of these experiments, comparing somewhat favorably with several competing learning models. We then extended the network model to deal with some attentional effects in human discrimination learning, wherein cue weight reflects attention to a cue. We further extended the model to include conjunctive features, enabling it to approximate classic results of the difficulty ordering of learning differing types of classifications. Despite the well-known limitations of one-layer network models, we nevertheless promote their use as benchmark models because of their explanatory power, simplicity, aesthetic grace, and approximation, in many circumstances, to multilayer network models. The successes of a simple model suggest greater accuracy of the LMS algorithm as against other learning rules, while its failures inform and constrain the class of more complex models needed to explain complex results.

We believe the adaptive network, or "connectionist," approach, briefly introduced in this issue, has considerable potential for solving some of the perennial problems of theoretical psychology. The articles in this issue, as well as the two volumes of papers edited by Rumelhart and McClelland (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986b), testify to the vitality and promise of this approach to theory in psychology.

Adaptive network models are being evaluated by several different methods. A first method, exemplified in the works of Sejnowski and Rosenberg (1986), Rumelhart and McClelland (1986a), Elman and McClelland (1988), and Hinton (1986), is to show how a complex adaptive network can learn structured information that organizes and underlays performances as diverse as speaking from written text, verb-tense production, contextual influences on phoneme recognition, the semantic components of family trees, and so on. Typically, the models required to produce such complex phenomena have several layers of hidden units and presume considerable structure in the input units and/or the output units (e.g., the NETtalk model of Sejnowski & Rosenberg, 1986). Such demonstrations are valuable and have a status something like "existence proofs" in mathematics or proofs in computer science that a program

The work presented here has benefited from the comments and insights of Andy Barto, Nelson Donegan, William K. Estes, Richard Golden, Daniel Kahneman, Doug Medin, James McClelland, Michael McClosky, Robert Nosofsky, David Parker, Misha Pavel, Paul Rosenbloom, David Rumelhart, Roger Shepard, Richard Sutton, Richard Thompson, and Amos Tversky. For their assistance with this research, we are grateful to Andrea Gallagher, Carrie Henderson, Anthony Henin, Van Henkle, Robert Kylberg, Gus Larsson, Susanna Lee, Naomi Schechter, and Jamie Uyehara. This work was partially supported by a research grant (MH-13950) to G. Bower from the National Institute of Mental Health, by Grant BNS-8618049 from the National Science Foundation, and by the Sloan Foundation.

0749-596X/88 $3.00

is correct or computationally sufficient. In such cases, one is satisfied if major regularities and salient phenomena (e.g., over regularizing irregular verbs, as in *go-ed*) are being roughly captured by the simulation.

A second method for evaluating such models, perhaps one more familiar to experimental psychologists, is to apply an adaptive network model to a familiar experimental paradigm that has been extensively studied and to see to what extent the model can account for some of the well-known results in that area, and perhaps predict detailed quantitative aspects of results of new experiments which may discriminate among alternative theories. This has been the traditional research strategy of many workers in cognitive psychology. That approach to testing network models was followed by McClelland and Rumelhart (1981) in their word perception model, by McClelland and Elman (1986) in their speech perception model, and by Dell (1986) in his model of errors in speech production. It is also the approach we have followed for the research we shall report. Because we are interested in human learning and categorization, those are the areas that have attracted our initial efforts. After we present some new experiments on classification learning and show how the network model handles them, we extend the model to handle several phenomena suggesting selective attention and hypothesis testing in concept learning.

In developing an adaptive network model for any specific learning task, three sets of assumptions are required.

- First, one must assume a particular "architecture" of the connected units: will it be only a feed-forward system, or one with recurrent (return) connections or within-layer connections? Will there be hidden units intermediate between input and output layers? If so, how many layers of hidden units? How many units and how will they be connected?
- Second, one must decide how to represent the learning materials within this architecture. In particular, what corresponds in the network to presentation of an experimental stimulus pattern? What metric over the set of "output units" corresponds to a behavioral response that one can measure in the experiment?
- Third, one must decide how learning occurs. By what algorithm or rule are the connection weights to be adjusted trial by trial to simulate adaptive learning in the network?

The many options in these decisions suggest that it is difficult to test the adaptive network framework in general. Rather, one can only test a specific realization of the framework. By noting the circumstances where it predicts accurately versus those where it has shortcomings, we can gather generalizations about which network assumptions and learning algorithms are generally adequate to explain results across broad ranges of experimental conditions.

Some History and Beginning Considerations

A long tradition of research in learning has been based on the idea that mechanisms of learning would be the same throughout mammalian species. This assumption of phyletic continuity underlay much of the research interest in conditioning in animals (rats, cats, monkeys, pigeons) and it justified the facile mixing of results from human and animal learning experiments in the writings of prominent "learning theorists" such as Edward Thorndike, Edwin Guthrie, Edward Tolman, and Clark Hull (see Bower and Hilgard, 1981). The early models of neural networks for learning (e.g., McCulloch & Pitts, 1943; Rashevsky, 1937) fell squarely within this tradition as did Rosenblatt's (1961) later "Perceptrons"; both approaches aimed to derive processes of complex learning from configurations and

elaborations of a small set of elementary associative processes that could be observed in lower animals.

About 25 years ago, however, the popularity of theories based on animal conditioning began to wane, whereas theories and studies of human memory underwent a major transformation as well as a growth spurt. The theoretical view of memory was revolutionized by ideas spun off from the metaphors of information processing and the mind-as-computer—notions of memory stores, selective coding, capacity limits, organization, labeled semantic networks, rule-based symbol manipulation, production systems, and so on. Such concepts and theories seemed to have more of the power needed to deal with the richness of language and its role in encoding, storing, and retrieving human memories. Consequently, the theories of associative conditioning were somewhat "left behind" in the enthusiasm to embrace the new perspective. However, the separation of these two major fields for studying learning has never been a happy one, and it has instigated periodic attempts at reconciliation, including recent trends towards "cognitive" theories of animal memory (e.g., Roitblatt, 1987; Wagner, 1981).

The current resurrection of adaptive networks, as models of complex abilities such as parsing, reading, and speech recognition and production (see, e.g., Cottrell, 1985; Waltz & Pollack, 1985), may be cause to renew the traditional interest in subsuming human and animal learning under one set of associative mechanisms. Given the voluminous studies of learning in animals alongside current attempts to model cognition with elementary associative processes, a reasonable tactic is to search for and exploit any correspondences which might exist between animal and human learning. This is the tactic we have taken in part of our research, best exemplified in Experiments 1, 2, and 3 of Gluck and Bower (in press) to be reviewed below.

Our initial interest was in finding some correspondences between associative principles underlying human and animal learning. As a start, we decided to use the Rescorla–Wagner model of associative learning (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). The Rescorla–Wagner model is one of the most widely accepted descriptions of conditioning in animals and is founded on a wide range of confirmatory results.[1] Desiring to relate the Rescorla–Wagner model to human learning and to test some of its unique predictions within that domain, our interest in adaptive networks for this purpose was aroused by a paper by Sutton and Barto (1981). They noted that the Rescorla–Wagner rule for association formation was a special case of the least-mean-squares (LMS) learning rule widely used in training adaptive networks. (The LMS rule is variously called the Widrow–Hoff rule, after its originators, or the delta rule because of its use of differences.)

For the previous year we had been carrying out experiments on human category learning, wherein subjects learned to classify patterns of stimulus features. Given

[1] Despite the many successes of the Rescorla–Wagner model, it does have several well-known limitations and shortcomings. First, it does not explain learned irrelevance of a cue that has first been randomly paired (uncorrelated) with an unconditioned stimulus (US). Conditioning in the former case is severely retarded, relative to a neutral cue, by that earlier learned irrelevance (see Baker & Mackintosh, 1977). Second, one cannot drive to zero strength a conditioned inhibitor (with $V = -\lambda$) by presenting it without the US—although the Rescorla–Wagner model says that that should happen (see Zimmer-Hart & Rescorla, 1974). Third, reducing the number of USs (shocks) per trial from two to one in the two-phase experiment causes unblocking and learning of the second, redundant cue, in contradiction to the model (Dickinson, Hall, & Mackintosh, 1976). Despite these limitations, the Rescorla–Wagner model has remained for the last 15 years as the most elegant and widely accepted model of the associative changes occurring during classical conditioning; the wealth of confirmed implications arising from this deceptively simple model has been substantial.

that interest and the Sutton–Barto observation, we were motivated to develop an adaptive network model for our simple form of human category learning and to test the Rescorla–Wagner learning rule in that context. Because of its popularity in recent work by Medin and others, we chose as our test situation a simulated medical diagnosis task; student subjects learned to classify hypothetical patients (described by one to four medical symptoms) into one of two disease categories, receiving feedback about the correct diagnosis after each decision. After describing the model for this task and three tests of the model's predictions, we will discuss how the model deals with attentional and/or hypotheses-testing behaviors during discrimination learning.

The Basic One-Layer Network

Figure 1 shows the simplest one-layer network possible to deal with the case of two output (disease) categories. The input nodes correspond to the four medical symptoms. Presentation of a simulated patient corresponds to activating those input units representing that patient's symptom pattern. For example, a patient with bleeding gums, a runny nose, and stomach cramps might correspond in the model to turning on input units s_1, s_3, and s_4. If a patient does not have a specific symptom, we assume that the corresponding input unit has zero activation. Mathematically, each patient p can be represented as a vector with four binary components (symptoms), where the jth component, x_{pj}, has value α_j or 0 according to whether patient p does or does not have symptom s_j.

Subjects understood that each patient had either of two diseases but not both. A network model with two mutually exclusive categories can be simplified to a network with one output node (as in Fig. 1); this one output node reflects the difference in strength of activation in favor of category (disease) 1 over category (disease) 2. In principle, one could add a bias term, which must be exceeded to get a preference for a particular response. A bias term can be introduced by adding a constant input node, corresponding to the context, which is always turned on with a $+\alpha_j$ input activation, and which is then adjusted by the LMS algorithm just like the other weights. In this paper, we have not used the bias term in any of the model's applications. In general, the model with a bias term does not fit our three initial experiments as well as the network model without a bias term.

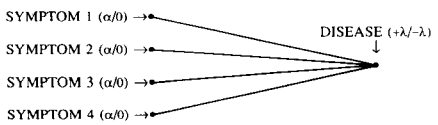

FIG. 1. A simple one-layer network with four input nodes and one output node.

Each input activation, x_{pj} is multiplied by its weight, w_j, and the sum appears on the output node, reflecting the extent to which the network favors disease 1 over disease 2 as the classification of patient p. Weight w_j reflects the strength of differential association between symptom j and disease 1 (versus 2). These weights are adjusted trial by trial by some learning rule so as to adapt to the correlations between symptoms and diseases.

Training of the network is carried out by presenting it with a series of patients (symptom patterns) and telling it which disease each has. We represent the training feedback on each trial as λ_1 if disease 1 is correct, and λ_2 if disease 2 is correct. For n input nodes, the learner is presumed to compare the calculated output $o_p = \Sigma_{k=1}^{n} x_{pk} w_k$ to the desired output (λ_1 or λ_2), and then to adjust each of the weights that contributed to the outcome so as to bring the calculated output closer to the desired output for that stimulus pattern. The change on the trial in the jth weight, w_j, is achieved according to the LMS rule:

$$\Delta w_j = \beta(\lambda_p - o_p)x_{pj}$$

$$= \beta\left(\lambda_p - \sum_{k=1}^{n} x_{pk}w_k\right)x_{pj}$$

$$= \begin{cases} \beta\left(\lambda_1 - \sum_{k=1}^{n} x_{pk}w_k\right)x_{pj} & \text{if disease 1} \\ \beta\left(\lambda_2 - \sum_{k=1}^{n} x_{pk}w_k\right)x_{pj} & \text{if disease 2} \end{cases} \quad [1]$$

Several features of Eq. [1] should be noted. First, no change in weights occurs for nonpresented stimuli, for which x_{pj} is zero. Second, the change in association of s_j to a disease on a given trial is smaller the closer the actual output is to that desired for the input pattern. Thus, a given training experience will cause a big change in the weights (associations) only if something "unexpected" happens. This also describes the conditions that produce blocking in classical conditioning (Kamin, 1969). Third, as noted by Sutton and Barto (1981), Eq. [1] is equivalent to the Rescorla–Wagner conditioning model wherein the strength V_i of cue i is altered according to $\Delta V_i = \beta(\lambda - \Sigma V_i)$, where the summation is taken over cues that are presented on the trial. Their Vs are equivalent to our ws; they suppress the x_{pj} notation, but these are implicit in the terms that appear in their equations. Fourth, in principle, the full network in Fig. 1 has 11 parameters—four input activation values (α_1, ..., α_4), four starting weights (w_1, ..., w_4 on trial 1), two training signals (λ_1 for disease 1 and λ_2 for disease 2), and the learning rate, β. But these parameters can be greatly reduced to essentially one, as we have done. We assume $\lambda_1 = -\lambda_2 = \lambda$ due to equal payoffs for the two categories, assume the initial w_js are zero, and assume the αs are equal due to equal salience of the four sensory inputs. Moreover, for given training contingencies, one can show that the asymptotic weights are unique up to a multiplicative factor of (λ/α). Consequently, we lose no generality by setting $\lambda = \alpha = 1$, as we have done in our calculations. All these restrictions leave only β as the unknown parameter. It determines the speed of learning, or equivalently, the speed of convergence to the least-mean-squares solution. For a given training sequence, Eq. [1] implies a learning curve (in ws and/or choice probability) that is usually negatively accelerated with a rate of approach to asymptote depending on β. The model can be tested by its fit to the learning curve observed under different conditions.

If the correlations between the stimulus patterns and outcomes are trained long enough, the LMS rule settles into a unique asymptotic set of weights independent of β (provided $\beta > 0$). This is termed the LMS solution to the problem. These weights determine the performance of the network. Therefore, the model can also be tested by its predictions about asymptotic performance after many learning trials. Of course, asymptotic predictions provide only an indirect indicator of the learning rule, but they are nonetheless useful when they differ from those of alternative models. The reader should note that these asymptotic predictions of the LMS network will often be parameter-free, determined only by the structure of the learning problem and the LMS rule.

Fifth, Eq. [1] can be viewed analytically as one of a class of functions that will optimize the accuracy of the model's predictions according to some criterion. Several criteria are plausible, including (1) minimizing the squared errors or discrepancies between the network's output and the desired output, averaged across all patterns, (2) minimizing the average percentage errors of the outputs, or (3) minimizing the expected cost of the errors. Equation [1] follows from adopting the first criterion, minimizing the squared errors of prediction (see Eq. [2]). Letting o_p and λ_p represent the actual output and desired output of the network for stimulus pattern (patient) p,

the expected value of the squared error, e, is

$$E(e) = E[\lambda_p - o_p)^2] = E\left[\left(\lambda_p - \sum_{k=1}^{n} x_{pk}w_k\right)^2\right]. \quad [2]$$

The expectation is taken over all types of input patterns, whereas the inner sum is over the n input nodes associated with pattern (patient) p. Weight w_js contribution to minimizing the expected squared error can be obtained by differentiating $E(e)$ with respect to w_j. In particular, using the chain rule,

$$\frac{\delta E(e)}{\delta w_j} = \frac{\delta E(e)}{\delta o_p}\frac{\delta o_p}{\delta w_j} = -2(\lambda_p - o_p)x_{pj}.$$

This derivative is proportional to the weight change dictated by the LMS rule, with a negative constant of proportionality. These weight changes carry out a "steepest descent" search for a minimum $E(e)$ in weight space (see Stone, 1986; Widrow & Hoff, 1960). The set of weights which produce a minimum $E(e)$ is called the LMS solution for that problem.

We use the error-correcting rule in Eq. [1] because it is simple, has historical precedents in Widrow and Hoff's (1960) adaptive learning networks, and importantly, is identical to the Rescorla–Wagner model of associative learning. Further, a number of useful theorems have been proven about the LMS rule in adaptive networks (Kohonen, 1977; Parker, 1985, 1986; Stone, 1986).

Many readers well recognize in Eq. [2] the ingredients of a *linear regression* analysis, wherein the input variables x_{pj} are treated as predictor variables for a criterion variable λ_p which is 1 or -1 for that pattern. Equation [2] provides a maximum likelihood, and least-squares, estimation of the w_is as regression weights. Like regression weights, the w_is reflect the correlation between the predictor variable, x_i, and the criterion, after correcting for intercorrelations among the predictor variables. Thus, Eq. [1] provides an iterative procedure to change the w_is trial by trial so as to converge asymptotically to the least-squares estimates of the regression weights. This correspondence implies that a linear-regression model would show many of the phenomena captured by the Rescorla–Wagner model when applied to a fixed set of stimulus–response pairings. The network model also bears an interesting relation to discriminant functions, the Brunswick lens model, logistic regression, and a Bayesian-inference model (see Slovic & Lichtenstein, 1971). However, to pursue these relationships would take us too far afield here.

Although the mathematical properties of the LMS learning rule have been well explored, we wish to address the question of whether the rule provides an empirically accurate account of how people learn. As a first step in evaluating the LMS rule as a component of human learning we began by exploring the accuracy of its predictions for asymptotic behavior of adults who have learned probabilistic classification ("discrimination") problems.

Experiment 1

Following procedures used by Medin and his associates, our subjects learned to classify stimulus patterns (hypothetical medical patients) into one of two disease categories. In the first experiment of Gluck and Bower (in press), students, serving as medical diagnosticians, read the medical charts of hypothetical patients, each described by the presence or absence of each of four symptoms. The symptoms were imperfect indicators, however, and had only probabilistic relations to the diseases. Thus, the situation was similar to paradigms for studying multiple-cue probability learning (Castellan, 1977) or the training of fuzzy, ill-defined categories (Medin & Smith,

1984). The student diagnostician classified each patient as having one or the other of two fictitious diseases, then received feedback regarding the correct diagnosis. During training, subjects eventually learned which symptoms were more or less diagnostic of each disease. At the end of training, subjects were asked to directly estimate the conditional probability that a patient who had symptom s_i (but with *no* information about his other symptoms) had one disease or the other. These probability estimates provide differentiating results for several models we considered.

Experiment 1 was designed to distinguish the predictions of the LMS network model (in Fig. 1 and Eq. [1]) from three popular, competing models of category learning: (1) *exemplar* models which presume that the learner stores all the exemplars of each category and then classifies a new instance according to its relative similarity to the stored exemplars of each category (e.g., Medin & Schaffer, 1978; Nosofsky, 1984), (2) *feature-frequency* models which presume that the learner stores relative frequencies of occurrence of cues within the categories and then classifies an instance according to the relative likelihood of its particular pattern of features arising from each of the categories (Franks & Bransford, 1971; Reed, 1972) and (3) *prototype* models which presume the learner abstracts the central tendency (average description) of each category and then classifies instances according to their similarity to this average prototype (e.g., Fried & Holyoak, 1984; Homa, Sterling, & Trepel, 1981).

Considering subjects' estimates of the probability of each disease given each symptom, these models make one of two predictions. Exemplar models predict that subjects would access all, or a random sample of the training exemplars which contained the specified symptom and note how often this symptom occurred with each disease. Thus, these models predict that subjects' estimates of the conditional probabilities will reflect the observed conditional symptom-to-category probabilities of the training sequence, a form of "probability matching." A pure prototype model which ignores the differing base rates of the diseases would predict that subjects' estimates of the probability of a disease given a symptom will reflect the closeness of that symptom to the value of that symptom in the prototypes of the two diseases. In the feature-frequency model, subjects are presumed to keep track of how often the symptoms (features) occur with each category (disease) and then transform these to conditional probability estimates. Thus, the feature-frequency model makes predictions identical to those of the exemplar model, viz., the estimates should reflect the normative conditional probabilities of the disease given each symptom as it was realized in the patterns shown in the training sequence. Estes (1986) provides a fuller description of the commonalities and differences between these models.

To creat predictions that differentiate the LMS network from the alternative models, we sought a learning task in which the *ordinal* relationships among the asymptotic weights (across diseases) predicted by the LMS model differs from the ordering predicted either by the objective posterior conditional probabilities of the categories given the features *or* by the relative likelihoods of the features given the categories. We discovered that one way to arrange such a situation was to *unbalance* the overall frequencies of the two diseases (their "base rates") so that one occurs far more often than the other. We will call these the common (C) and rare (R) diseases.

In Experiment 1, patients with the common disease were presented three times as frequently as patients with the rare disease. Patients' symptoms were selected so that the probability of each of the four symptoms occurring in patients suffering

from each of the two diseases was that shown in Fig. 2A. The lower numbered symptoms (different for every subject) were more typical for the rare disease while the higher numbered symptoms were more typical of the common disease.

Using the base rates of $P(R) = .25$ and $P(C) = .75$ and the probabilities in Fig. 2A, Bayes' theorem provides the conditional probability of the two diseases given the four symptoms considered separately (see Fig. 2B). For any single symptom the Bayesian probability of the rare disease was always less than or equal to the probability of the common disease.

Direct probability estimates. After 250 training trials of predicting diseases and receiving feedback, subjects were finally asked to estimate directly the probability that a patient exhibiting a particular symptom was suffering from the rare disease. They were explicitly told that information about presence versus absence of the other symptoms was to be considered unavailable when they made these judgments, and later questioning indicated their correct understanding of this point.

Figure 3 shows both the actual probabilities in the training patterns as well as the probability-matching behavior predicted by exemplar-storage and feature-frequency models. But the LMS rule predicts that following training, subjects' estimates of the probability differences will follow a different pattern, reflecting the underlying strengths of the feature-to-category associative connections. These asymptotic connection weights can be calculated from Eq. [1] by deriving equations for the expected trial-by-trial weight change in each of the feature-to-category connections, setting these expected changes to zero at asymptote, and solving the resulting four simultaneous equations in four variables. As noted earlier, the asymptotic connection weights depend only on the reinforcement probabilities in Fig. 2B and not on the learning rate, β.

The expected asymptotic association strengths turn out to be .430, $-.043$, and $-.306$, and $-.771$ for w_1 through w_4, respectively; these are plotted in Fig. 3. We assume that these differential association

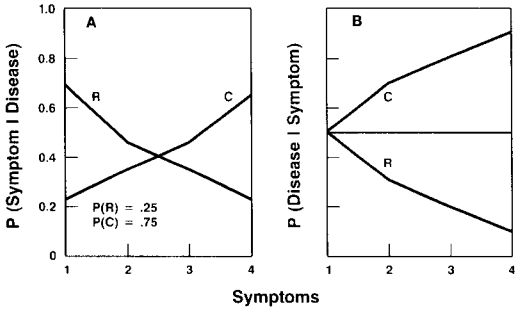

FIG. 2. Experiment 1 design: (A) The probabilities of each of the four symptoms occurring in patients suffering from each of the two diseases. The lower numbered symptoms were more typical of the rare disease while the higher numbered symptoms were more typical of the common disease. (B) The conditional probabilities of each of the two diseases given the presence of each of the symptoms computed from (A) using the base rates and Bayes' theorem.

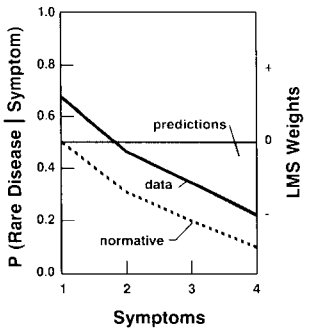

FIG. 3. Results and predictions for Experiment 1. The normative probabilities of the rare disease given each of the symptoms are shown as a dashed line (with the scale along the left vertical axis). These also correspond to the predictions of exemplar and feature frequency learning models. The predictions of the LMS rule, based on asymptotic levels of associations, are shown as shaded areas above or below the middle axis to indicate that they are to be interpreted relative to the scale on the right. These predictions are unique to within a scalar multiple. Hence, the critical aspect of the predictions is the relative degree to which they are either above or below the zero line, corresponding to a prediction of .5 on the left scale. The observed means of subjects' estimates of $P(R|s_i)$ are shown as a solid line (using the scale along the left vertical axis).

weights are monotonically related to the direct estimates that subjects make for each symptom–disease conditional probability. In particular, the direct estimate of the probability of disease R should be above, equal to, or below .50 accordingly as the theoretical w_i is above, equal to, or below zero, respectively. (Recall, the ws are differential weights for one category versus the other.)

Our initial comparisons will use only these ordinal properties of the weights of the different symptoms. The relevant ordinal comparisons are graphed in Fig. 3, showing the objective posterior conditional probabilities and the theoretical associative weights along with the data. The most striking difference between the objective probability measures in Fig. 3 and the theoretical associative weights in Fig. 3 occurs for symptom 1 (denoted s_1). This symptom was paired as often with the rare disease as with common disease; hence, the conditional probabilities were objectively .5. However, the LMS rule predicts that s_1 will be somewhat more associated with the rare disease than the common disease, i.e., that $w_1 > 0$.

To see the basis for this prediction, we note that the asymptotic *symptom→disease* weight reflects the degree to which a symptom has been a valid predictor of a disease *relative* to the predictive value of other symptoms that might be present at the same time. Although s_1 has the same predictive value for the two diseases, it is a relatively better predictor for the rare disease than are any of the other symptoms. On rare disease trials in which s_1 occurred, the other symptoms were less likely to be present; and if they were, they were more strongly associated with the other disease. Hence, according to Eq. [1], w_1 was pushed overall more toward +1, indicating the rare disease, than toward −1, indicating the common disease.

Turning now to the data in Fig. 3, comparison of the actual with the estimated conditional probabilities indicated that while subjects correctly learned the relative strengths of the conditional probabilities within a particular disease category, they appreciably *overestimated* the conditional probability of the rare disease given each of the symptoms. The data for s_1 are critical for distinguishing between the models. The data indicate that subjects believed that patients with symptom s_1 were significantly more likely to be suffering from the rare disease than from the common disease. This result is exactly as predicted by the LMS network model.

Predicting choices to patterns. We have also used the LMS model to predict successfully the observed choice proportions given each of the 15 possible symptom patterns. For each symptom pattern, we obtained the summed weights and converted that activation into a probability of assigning that patient to the rare disease category. For this conversion, we used the logistic output function, viz.,

$$P(R|\text{pattern}) = \frac{1}{1 + e^{-\theta \sum_{k=1}^{4} w_k s_k}}. \qquad [3]$$

The logistic output function is a close approximation to the normal integral function of the difference between the strengths of the two response alternatives, which was the choice rule proposed long ago by Thurstone (1927) and Hull (1943). It is also the oft-used output function used in connectionist networks (see McClelland and Rumelhart, 1986).

Figure 4 shows the fit of the logistic to the 15 patterns' choice proportions estimated from the last 50 training trials. The best-fitting value of θ, as estimated by least squares from the observed proportions, is 3.20. The correlation between observed and predicted choice proportions is .94 with an average absolute discrepancy of .07, suggesting a reasonable fit of the model to these data. However, the χ^2 is 46.2(14), p

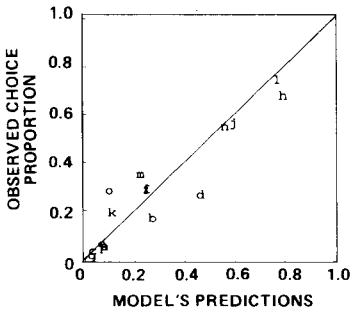

FIG. 4. Predicted vs observed proportions of choices of the rare disease given 15 different symptom patterns, labeled *a* through *o*, from Experiment 1 of Gluck and Bower (1987).

$< .001$, suggesting significant deviations from the model. We would point out that many of the choice proportions are based on relatively small samples, so have large standard errors.

An interesting fact is that the observed and predicted choice proportions are close to the objective probabilities that any particular pattern has been paired with the rare disease. This is a form of "probability matching" at the level of entire patterns of symptoms and it corresponds to predictions of Estes' pattern model (Estes, 1960). We note that such probability matching violates the "optimal" decision rule which dictates that one should always assign the rare disease to any patients whose sum of weighted symptoms exceeds zero, and otherwise assign the common disease to them. Of course, this suboptimality has often been commented upon in the literature on probability matching (e.g., Estes, 1972).

Setting aside these choice proportions, the primary focus of our ordinal analyses was on the subjects' direct estimates of the conditional probabilities of each disease given each symptom singly. The relevant differences between these probability estimates, shown in Fig. 3, confirm the predictions of the LMS rule. Importantly, subjects conformed to prediction in believing that symptom s_1 was a stronger predictor of the rare disease than of the common disease, although objectively the two diseases were equally likely whenever symptom s_1 appeared. Subjects behaved as though they were neglecting the higher base rate of the common disease.

This result suggests that our learners fell prey to a common form of "base rate neglect"; in making predictions, they erroneously judged that the presence of a symptom (s_1) highly representative of the rare disease was strong evidence for diagnosing the rare as opposed to the common disease. This result brings to mind several results in research on probability judgments: people consistently overestimate the degree to which evidence that is representative or typical of a rare event is actually predictive of it (Kahneman & Tversky, 1973). Most studies demonstrating such neglect of base rate in judgments have used natural categories with familiar prototypes (e.g., feminists or engineers), and base rate information has generally been presented to subjects as abstract numerical information (Tversky & Kahneman, 1982). Our demonstration of base rate neglect has arisen where information about categories and base rates was learned by subjects from examples. Of course, there is no assurance that the two forms of base rate neglect are generated by similar causal mechanisms.

Our results for direct probability estimates raise one or another problem for the three competing models of category learning introduced earlier. In brief, the problem is that these models fail to predict the subjectively high diagnosticity of symptom s_1. Rather, they expect that direct estimates of the conditional probability of the rare disease given symptom s_1 will match the objective probability (of .50)—and that was not so. Although the network model predicts more accurately in this instance, the other models, especially Medin's context model, have dealt successfully with a range of results that the simple network model would have difficulty explaining (e.g., Medin, Altom,

Edelson, & Freko, 1982; Medin & Schwanenflugel, 1981). Such results arise whenever the correct category depends on the joint value of two or more stimulus dimensions in a nonlinear manner. We will return later to this topic.

Experiment 2

In the second experiment of Gluck and Bower (in press), we tested a basic property of the LMS model, namely, that different cues *compete* to be the more valid predictor of an outcome. To the extent that a stimulus cue is redundant with a stronger or more valid cue in predicting an outcome, the model and Eq. [1] expect that that cue's associative strength will be greatly attenuated. Thus, in our category-learning paradigm, we expected to attenuate the apparent diagnosticity of symptom 1 (found in Experiment 1) by making some of the other symptoms truly reliable and strong predictors of the rare disease. To test this prediction, Experiment 2 was designed identically to Experiment 1 in all respects except that symptoms 2 (s_2) and 3 (s_3) were changed to be more valid predictors of the rare disease and common diseases, respectively. In particular, the probability of s_2 given the rare disease was set at .90 and the probability of s_3 given the common disease was set at .90. This outcome is predicted as well by the equivalent linear regression model, viz., the regression weight of one variable is reduced accordingly as it is partially correlated with a second, more valid predictor.

As predicted by the LMS rule, the results obtained in Experiment 2 showed that the apparent diagnosticity of symptom 1 for the rare disease was significantly attenuated by introducing greater validity for symptom 2 toward the rare disease. The direct estimate of the conditional probabilities of the rare disease given symptoms s_1 (without information of other symptoms) dropped from .67 (in Experiment 1) down to .59 in Experiment 2. This was a significant reduction, as the model predicted. It was as though symptom 1 "lost its punch," or its claims on the subject's attention, when it was put into competition with a truly diagnostic symptom for the rare disease. It is this *competitive* nature of cue–outcome association learning that gives the LMS rule its distinctive advantage over other learning rules. The rule implies that people do not learn sets of *independent* cue–outcome associations, but rather the relative diagnosticities of the various cues.

In further testing, the model was somewhat close in predicting the asymptotic choice proportions for each of the symptom patterns: using Eq. [3] with $\theta = 4.6$, the predictions yielded a fit to the observed proportions with a correlation of .97 and an average discrepancy of .09 (χ^2 (14) $= 53$, $p < .001$). The observed choice proportions were close once again to the objective probability-matching values.

Experiment 3

We noted in both Experiments 1 and 2 that the observed choice proportions for symptom patterns were close to the objective probabilities, as predicted by Estes' pattern model.[2] The correspondence was sufficiently close as to make the pattern model a strong competitor to the LMS network model. Consequently, Experiment 3 of Gluck and Bower (in press) sought to find evidence that would discriminate between the two theories. One crucial point we noted is that the pattern model treats symptom patterns as unanalyzable wholes (configural Gestalts), each one of which differs equally from the other patterns. (The "mixed" model of Atkinson & Estes, 1963, was proposed to deal with generaliza-

[2] Medin's context model can also predict probability matching to patterns in case the similarity parameters are set to zero; i.e., a test pattern matches only its own representations in memory and any mismatches would be discarded. Of course, the model would then fail to generalize to novel patterns.

tion from patterns to stimulus components.) As a consequence of this unanalyzability assumption, the pattern model treats presence versus absence of a single symptom within a pattern as equivalent to presence of a symptom versus its complement (or opponent) symptom in the pattern. But suppose instead of presence vs absence of the four symptoms in Experiment 1, that we substituted presence of the symptom versus its opposite symptom, such as runny nose vs stuffy nose.[3] Each patient would then be described by exactly four symptoms, by selecting one from each of the four opponent pairs of symptom values. Such an experiment would still have $2^4 = 16$ distinct stimulus patterns. If these were to have the same correlations with the diseases as did the patterns in Experiment 1, then the pattern model has no grounds for expecting a difference in outcome for the two cases.

The theoretical situation is different for the LMS network model. In fact, there are at least two different ways to represent the opponent-symptom situation within the network framework, and they make significantly different predictions. The two network representations are depicted in Fig. 5. Network A represents each opponent-symptom pair as a single input node which receives activation of $+\alpha$ or $-\alpha$ depending on which member of the opponent pair is presented on a given trial (patient). Alternatively, network B represents each symptom s_i and its opponent of the pair, s_i^*, as two distinct input nodes, one, and only one, of which is activated for each patient; this comprises eight input nodes. Notice that network A has a built-in negative correlation between any symptom and its opponent, whereas network B is silent on this issue. Insofar as it takes a stand on the issue, the linear regression model would

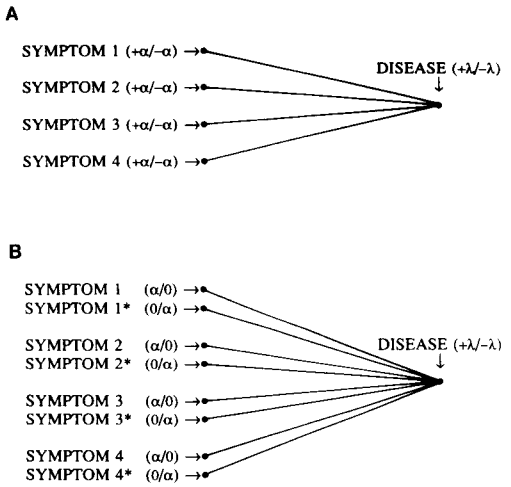

FIG. 5. (A) A four-component network for classifying the stimuli from Experiment 3 which represents each opponent-symptom pair as a single input node which receives activation of $+\alpha$ or $-\alpha$ depending on which member of the opponent pair is presented on a given trial (patient). (B) An eight-component network for classifying the stimuli from Experiment 3 which represents each symptom s_i and its opponent of the pair, s_i^*, as two distinct input nodes, one (and only one) of which is activated for each patient.

treat the experiment in terms of network A. If it tried to represent the experiment in terms of an eight-variable linear regression equation similar to network B, it would compute a perfect negative correlation between each s_i and s_i^*. Hence, the eight-variable regression model would collapse to a four-variable model with input values of $+1/-1$ rather than $+1/0$.

Both of these network representations are plausible and, indeed, correspond to different stances in the adaptive network literature. One obvious difference is that network A implies strong symmetry in activation for a given pattern and its complement (obtained by using the alternate values in the pattern). For example, if pattern $s_1 - s_2^* - s_3^* - s_4$ yields output activation G, then its complementary pattern $s_1^* - s_2 - s_3 - s_4^*$ will yield output $-G$. Therefore, the eight complementary pairs of patterns provide strong tests of this symmetry prediction from network A. On the other hand,

[3] Tversky (1977) called these substitutive features (e.g., color of eyes) rather than *additive* features (presence/absence of glasses).

network B, treating the eight symptoms as distinct, makes no such strong predictions. In viewing this experiment in terms of either network, we note that a symptom never can occur alone; rather, it always occurs in company with three other symptoms, so by the LMS rule it cannot acquire dominating associations. The competitive nature of the LMS rule thus leads to different outcomes for Experiment 3 than for Experiment 1 (under both network models).

Turning first to the models' predictions, Fig. 6 shows the predicted weights for the eight-component model. Also shown are the probability-matching predictions of Estes' pattern model, and, for comparison, the association weights predicted by the LMS model for Experiment 1.

Several features of these predictions may be noted. First, the pattern model's predictions of probability estimates for the positive symptoms s_1, s_2, s_3, and s_4 in Experiment 3 (see Fig. 6) are identical to what they were in Experiment 1 (see fig. 3). Second, the predictions of the LMS model of the subjects' direct estimates of $P(R|s_i)$ for symptoms s_1 through s_4 (viz., the w_is) differ mainly by being less extreme (i.e., closer to .5) in Experiment 3 than they had been in Experiment 1. Thus, compared to Experiment 1, symptom s_1 in Experiment 3 should appear to be somewhat less diagnostic of the rare disease, while symptoms s_3 and s_4 will appear less diagnostic of the common disease. Third, these two models differ in their predictions for symptom s_4^*. The pattern model implies that s_4 will be somewhat associated with the common disease, whereas the LMS model predicts that s_4^* will be more associated with the rare disease.

These differing predictions led us to conduct Experiment 3. As noted, the statistical design of Experiment 1 (see Fig. 2) was repeated with the exception that each present/absent symptom (e.g., fever or not) was replaced by two mutually exclusive features (e.g., stuffy/runny nose), one of which was always present for each patient. One slight difference was that Experiment 1 did not present a pattern in which a patient had none of the four symptoms, whereas the analogous case, of all complimentary features, was presented in Experiment 3. Thirty-six college student subjects

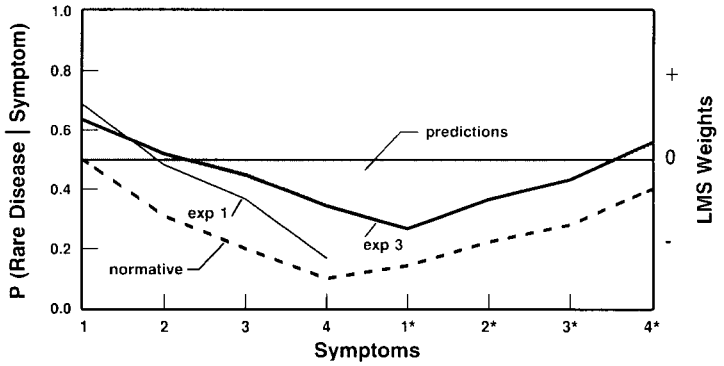

FIG. 6. Results and predictions for Experiment 3. The normative probability of the rare disease given each of the symptoms is shown as a dashed line, with the scale along the left vertical axis. These also correspond to the predictions of exemplar and feature frequency learning models. The predictions of the LMS rule are shown as shaded areas above or below the middle axis to be interpreted relative to the scale on the right. The observed data from subjects' estimates of $P(R|s_i)$ are shown as a solid line, using the scale along the left vertical axis. The data from Experiment 1 are also shown for comparison. Note that the normative conditional probabilities for symptoms 1 through 4 were unchanged from Experiment 1 to Experiment 3.

classified 250 patients, receiving feedback on each. After training, the subjects' estimated the conditional probability of each disease given each of the eight single symptoms (four mutually exclusive pairs).

Figure 6 shows the objective conditional probability estimates for the rare disease for each of the eight symptoms. The data of Experiment 1 are also shown for comparison. A number of conclusions may be drawn from these findings.

First, we consider the implications for the four-component model. As noted earlier, network A which represents the pairs of opponent symptoms by $+\alpha/-\alpha$ on four input nodes predicts that judgments of $P(R|s_i)$ should be symmetric to $P(R|s_i)$ around .5. The data, however, disconfirm this strong prediction. To take just one example, subjects' estimates of $P(R|s_3)$ and $P(R|s_3^*)$ were both significantly below .5, whereas one should be above .50; thus, this pair and several others violate the symmetry prediction of network A.

Also model A predicts for any symptom pattern T and its complementary pattern T^* that the choice probabilities are constrained by the relation $P(R|T) + P(R|T^*) = 1$. These predictions contrast sharply with the normative values of these probabilities. Contrary to this model's predictions, the observed choice proportions were very close to the normative probabilities. Thus, these observations permit the rejection of network model A.

Turning to the eight-component network model B, here the LMS rule predicts that symptom s_1 will still be considered more diagnostic of the rare disease, in contrast to the .5 value expected by the probability-matching models. The data support the LMS prediction. Second, model B predicted that the opponent symptom s_4^* would be considered diagnostic of the rare disease; the results support this outcome as against the probability-matching prediction of the pattern model (see Fig. 3). Third, in Experiment 3, we expected the probability estimates for symptoms 1, 3, and 4 to "shrink" toward .5, reflecting less competitive dominance than in Experiment 1. This too was observed.

Overall, then, the probability ratings appear to support the eight-component LMS model B as against the four-component LMS model A and Estes' pattern model. Both the eight-component LMS model and the pattern model predict a V-shaped graph over the eight symptoms in Fig. 3, but the LMS model predicts a V that rises above the .5 baseline at both ends. Also, the pattern model expects the probability estimates in Experiment 3 for s_1, s_2, s_3, and s_4 to be identical to those in Experiment 1. But in fact the two profiles for these symptoms deviate significantly from one another. The results here are not simply regression to chance due to slow learning with more cues in Experiment 3; rather, the alternative models do not account for the salient results predicted by the LMS model, namely, the significant *overexpectation* of the rare disease given s_1 and s_4^*.

Discussion of Experiments

The results of these three experiments provide preliminary converging evidence that the LMS rule is more general than formerly believed and is not limited to animal learning or to unobservable subcognitive changes in associations. Medin and Edelson (in press) have also obtained evidence that co-occurring cues compete for associations according to their relative validity. One limitation of the Gluck and Bower (in press) experiments is that they focus only on the asymptotic performance of subjects after extensive training. Thus, these studies only indirectly evaluate the learning process specified by Eq. [1]. More precisely, the results provide evidence for asymptotic LMS optimization, not for the learning rule which converges to that asymptote. The advantage of focusing on the asymptotic performance was that the model provided parameter-free predictions

about behavior which contrasted with the predictions of popular alternative models. Clearly, however, further work must be done to more directly evaluate the learning process itself.[4] MacMillan's (1987) study of zero-contingency cues is a very promising step in that direction, providing more direct evidence for the LMS learning rule.

Exploring farther afield, we have reexamined a few other phenomena of classification in light of the one-layer network model. As one instance, this model (as does Medin's & Hintzman's (1986) MINERVA 2) appears to handle much of the data previously cited in support of the prototype model. In particular, the model explains the graded typicality effects found in those classification experiments in which category exemplars are distortions of a central prototype (e.g., Franks & Bransford, 1971; Posner & Keele, 1968; Reed, 1972). Provided that a wide range of distortions of the prototype has been shown during trailing, the network will estimate weights that will cause the full prototype to be rated as most typical of the category, even though it was not shown during training. The adjusting weights are also sensitive to the frequency of particular exemplars within a category, so that items near the more-frequent exemplars come to be classified more accurately, as reported in the literature (e.g., Neumann, 1974; Nosofsky, 1988).

We have hardly begun our examination of the extensive literature on category learning using this simple network model. It is clear already that it will encounter difficulties fitting data obtained when people learn discriminations based on correlated cues (Medin et al., 1982) or learning categories based on nonlinear combinations of cues (Medin & Schwanenflugel, 1981). We shall return to this point in the next section. We are nonetheless encouraged not only that the LMS rule of adaptive network theories which fits these human learning data links with the Rescorla–Wagner model of conditioning, but that it also implies the phenomenon of base rate neglect which has proven to be a robust phenomenon in the literature of judgment and decision. We turn now to examine several phenomena which were once considered beyond the scope of simple conditioning models, namely, *selective attention* and *hypothesis testing* during category learning. We first extend the network model to deal with some attentional effects in category learning, wherein cue weight reflects attention to a cue. Second, we extend the model to include conjunctive features and show how this enables it to fit classic results on the difficulty ordering of learning differing types of classifications.

Attentional Phenomena in Learning

One deficiency of the association-based learning theories of the early 1960s was their difficulty in accounting for attentional phenomena in discrimination learning (e.g., Sutherland & Mackintosh, 1971). But some progress has occurred in extending conditioning models to account for attentional phenomena in animal learning (Rescorla & Wagner, 1972). It may be timely, therefore, to reexamine the relevance of conditioning models for attentional phenomena in human learning.

Two broad classes of models have been proposed in the conditioning literature to account for attentional phenomena (Rescorla & Holland, 1982). The first, and oldest, class of models emphasized variations in the processing of the stimulus cues (the CSs) due to a *limited attentional capacity*. As examples, the theories of Sutherland and Mackintosh (1971) and Trabasso and Bower (1968) both suggested that multiple CSs compete for a share of the organism's attention. For instance, in the

[4] W. K. Estes (personal communication, 1987) has reported some success in fitting the LMS rule to the learning curves of choices produced by subjects being tested with specific sequences of classified patterns.

blocking experiment of Kamin (1969), pretraining to the CS_1 stimulus was presumed to cause the animal to attend exclusively to this stimulus during the compound CS_1 + CS_2 training, effectively rendering the CS_2 stimulus nonfunctional; thus, no new association to CS_2 could occur during training on the compound. Related theories by Mackintosh (1975) and Pearce and Hall (1980) similarly emphasized variations in the processing of cues due to their history of informativeness.

An alternative class of models accounts for attentional phenomena such as blocking by postulating variations in associative learning due to differential processing (or impact) of the unconditioned stimulus (US). This view, originally proposed by Kamin (1969) and formalized by Rescorla and Wagner (1972), claims that the effectiveness of the US for promoting associative learning varies with the degree to which the US is surprising or unanticipated. In particular, the degree to which an outcome causes a stimulus element to become associated to it on a trial is proportional to the degree to which that outcome is surprising (unexpected) given *all* the stimulus elements present on that trial (Eq. [1] has this property). In such a model, a cue that acquires a strong association to an important outcome acts like a salient cue that attracts attention insofar as it largely controls the subject's behavior.

Within the animal conditioning literature, investigators have long debated the relative merits of these two models (see the review by Rescorla & Holland, 1982). In the literature on attention in human learning, however, all the models use the first approach, that of limited attentional capacity. For instance, models of attention in human learning, such as the earlier models of Zeaman and House (1963) or Trabasso and Bower (1968), or Nosofsky's (1986) recent generalization of the Medin and Schaffer (1978) context model, employ limited capacity assumptions.

In light of our evidence for the LMS rule in human learning, it is instructive to examine some conditioning phenomena which have strongly bolstered the differential association model of Rescorla and Wagner in contrast to the selective attention theories. Two phenomena in animal conditioning, namely, overexpectation and supernormal conditioning, provide discriminating analogs to ponder for human learning. First, in the overexpectation paradigm, two stimuli, CS_1 and CS_2, are first separately conditioned to the same unconditioned stimulus, and are then presented simultaneously in a compound CS_1 + CS_2 paired with the *US*. The LMS model implies that the initial training to the individual cues will drive both w_1 and w_2 to their asymptotic strength of λ, so that the compound association strength w_1 + w_2 will equal 2λ asymptotically.[5] On such compound trials, because the US that occurs is overpredicted the association strengths of the two cues are expected to actually *decrease* rather than increase. Furthermore, if a neutral stimulus, CS_3, were to be presented along with the CS_1 + CS_2 compound, all followed by the overexpected US, CS_3 should become *inhibitory*, according to Eq. [1]. These startling predictions have been confirmed by Rescorla and Wagner (1972) and Kremer (1978). Because limited attentional capacity models cannot predict that the old stimuli will lose stength during compound training nor that the novel stimulus will gain inhibitory powers, the results provide quite discriminating evidence in favor of the LMS approach.

A second discriminating result is supernormal conditioning which arises whenever a compound CS_1 + CS_2 stimulus paired with a US includes a stimulus, CS_1, which

[5] Rescorla and Wagner use V_i to denote the associative strength between CS_i and a US. But their Vs are formally equivalent to our weights, the ws, so we will continue to use the weight terminology.

has previously been conditioned as an inhibitor (predicting absence of a US). To be specific, suppose the prior inhibitory conditioning of CS_1 causes w_1 to equal $-\lambda$, whereas CS_2 begins as a neutral cue with $w_2 = 0$. The LMS rule implies that training with only the compound $CS_1 + CS_2$ paired with the US will change the two weights equally, so that their difference will always equal their initial difference, λ. This means that w_1 will increase more rapidly than it would have in a control condition where both weights were initially equal. Moreover, if a conditioned inhibitor (CS_1) is presented in compound with a former conditioned excitatory stimulus (CS_3) and the compound is reinforced, then the excitatory CS_3 will acquire even greater strength. This occurs because the strength of the compound ($w_1 + w_3$) will be less than the asymptote λ, so that the LMS rule will strengthen CS_3 even more than it was at the end of its initial training. Rescorla (1971) and Wagner (1971) confirmed that both these training procedures produce supernormal conditioning of the excitatory stimulus component. Limited attention models do not provide as direct an explanation for these results as does the LMS model.

A third major attraction of the Rescorla–Wagner model has been its ability to account parsimoniously for both excitatory and inhibitory conditioning (Wagner & Rescorla, 1972). For example, if in a conditioned inhibition experiment CS_1–US trials are intermixed with $CS_1 + CS_2$ – *no US* trials, the Rescorla–Wagner model correctly predicts that such training should result in CS_2 acquiring inhibitory properties. Moreover, the presence of a familiar inhibitory cue will *block* the acquisition of inhibitory power by a new second cue which is paired with the first one in predicting absence of the US. Limited capacity attention models, however, are silent about such issues, and so suffer in comparison to the Rescorla–Wagner account.

Attention Optimization and the LMS Rule

Despite such evidence from the animal learning literature favoring the LMS rule over the selective attention theory, the latter holds a favored position in the literature on human discrimination learning. A recent example is Nosofsky's (1984, 1986) very elegant model of how subjects distribute their attention among the available cues in the task. So, as a first step in evaluating the LMS model as an attention-like mechanism in human learning, we tried to fit some of the same results as did Nosofsky with his model.

Nosofsky's model calculates how a subject should best distribute his attentional capacity over the available cues so as to optimize his classificatory performance. One implication is that the subject should pay more attention to more valid (predictive) cues and less to irrelevant cues. In the model, differential attention to a stimulus dimension (e.g., dark vs light, circle vs ellipse) is reflected by more or less generalization between the values on that dimension. Values along unattended dimensions are less noticed and thus more similar. These intradimension similarity values are then used along with the Medin and Schaffer (1978) exemplar model to calculate response proportions to stimulus patterns in a variety of discrimination-and-transfer experiments (see Nosofsky, 1984, 1986). In this respect, the model is commendably accurate.

An acknowledged shortcoming of Nosofsky's model, however, is that it is static. It simply assumes that subjects during the course of training adopt an optimal distribution of attention (over the stimulus dimensions), but the model does not specify a step-by-step *learning mechanism* that could arrive at this optimal distribution as a result of training. How do subjects divine the optimal allocation of attention? We asked whether the asymptotic performance of the LMS rule might turn out to be nearly

equivalent to the optimum performance dictated by Nosofsky's model of attention. Let us examine a specific case.

Nosofsky (1984) analyzed an experiment by Medin, Dewey, and Murphy (1983) in which subjects learned to classify headshot pictures of women that varied along four binary dimensions: hair color (light or dark), hair length (long or short), shirt color (light or dark), smiling face (open or closed mouth). The category structure used in this experiment is presented in Table 1 where 1 and 2 denote the two values on each dimension, labeled at the top of the columns. During training, pictures A1 to A5 were assigned to category A, while pictures B1 to B4 were assigned to category B. Table 1 also shows seven novel test patterns (N1 to N7) which were presented in a testing session after training. Models may be tested by their ability to account for the probabilities of different category responses to these test patterns.

Nosofsky's model characterized the attention to each dimension by a similarity parameter which reflects the weight of that dimension's contribution to the classification judgments given to the test patterns. A highly relevant or informative dimension (such as dimension 1 in Table 1) would be reflected in a high attention weight. These weights are parameters that can either be estimated by fitting subjects' classification performance or they can be derived theoretically from an optimization scheme as Nosofsky did.

We have used the LMS rule to simulate the experiment of Medin et al. by training simulated Monte Carlo subjects to asymptote on the stimuli in Table 1. Figure 7 shows the theoretical parameters estimated from the Medin et al. data. The three lines in the figure are the observed attention weights as estimated from the data (based on transformations of the similarity parameters in Medin's context model), the predictions from Nosofsky's (1984) attention optimization model, and the normalized association strengths derived from the LMS network model for this experiment. Recall, the magnitudes of these values correspond to the asymptotic importance of the dimensions in the categorization task, i.e., the informativeness of the dimension for distinguishing the categories and/or controlling the response. These predictions are com-

TABLE 1
CATEGORY STRUCTURE FROM MEDIN ET AL. (1983)

		Dimension			
	Exemplar	1	2	3	4
Category A	A1	1	1	1	2
	A2	1	2	1	2
	A3	1	2	1	1
	A4	1	1	2	1
	A5	2	1	1	1
Category B	B1	1	1	2	2
	B2	2	1	1	2
	B3	2	2	2	1
	B4	2	2	2	2
Transfer test patterns	N1	1	2	2	1
	N2	1	2	2	2
	N3	1	1	1	1
	N4	2	2	1	2
	N5	2	1	2	1
	N6	2	2	1	1
	N7	2	1	2	2

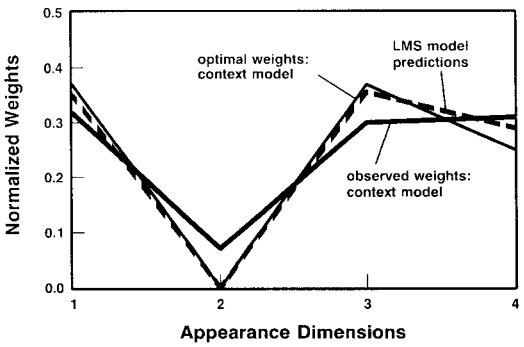

FIG. 7. Analyses of stimuli and data from Medin et al. (1983) showing the observed attention weights as estimated from the data (based on transformations of the similarity parameters in Medin's context model), the predictions from Nosofsky's (1984) attention optimization model, and the normalized association strengths derived from the LMS rule. (Adapted from Nosofsky, 1984, with permission).

pletely parameter-free for the LMS model since they are a function solely of the structure of the way the examples in Table 1 are classified.

Figure 7 shows that stimulus dimension 2 was the least important whereas dimensions 1, 3, and 4 were about equal in importance for both the data and the two models. Nosofsky's attention optimization hypothesis and the LMS rule deliver almost identical predictions. Not only do both provide close predictions of the data, but the predictions of the two models, in each dimension, differ from the data in the same direction. Thus, both models appear to be capturing the same regularities and characterizing the informativeness of the stimulus dimensions similarly.

To further compare the two models, we analyzed the models' predictions of choice probabilities for the nine training exemplars and the seven test exemplars (see Table 1). Figure 8B shows a scatterplot of the observed choice probabilities (from the experiment of Medin et al., 1983) compared to the predictions of the context model. Figure 8A plots the same data against the LMS rule. Predictions of choice probabilities from the LMS rule were made by entering the asymptotic association strengths (toward the two categories) shown in Fig. 7 into Eq. [3] with $\theta = 1.5$. Predictions from the context model, as reported in Medin et al. (1983), were made by fitting four similarity parameters, one for each dimension of appearance of the women's pictures. As shown by the points clustering near the diagonal, both the LMS and Medin models do exceptionally well in predicting subjects' choice probabilities for the different patterns. The predictions of the models are quite similar to one another for the 16 individual patterns (see Table 1).

Given the similar predictions of the LMS and Nosofsky's models, an interesting question is whether there is some formal relationship between the two. Recently, D. Rumelhart and R. Golden (personal communication, 1987) have shown that under certain conditions the LMS model with the logistic output function closely approximates Nosofsky's attention optimization model. The approximation is good whenever it is possible to find a set of attention

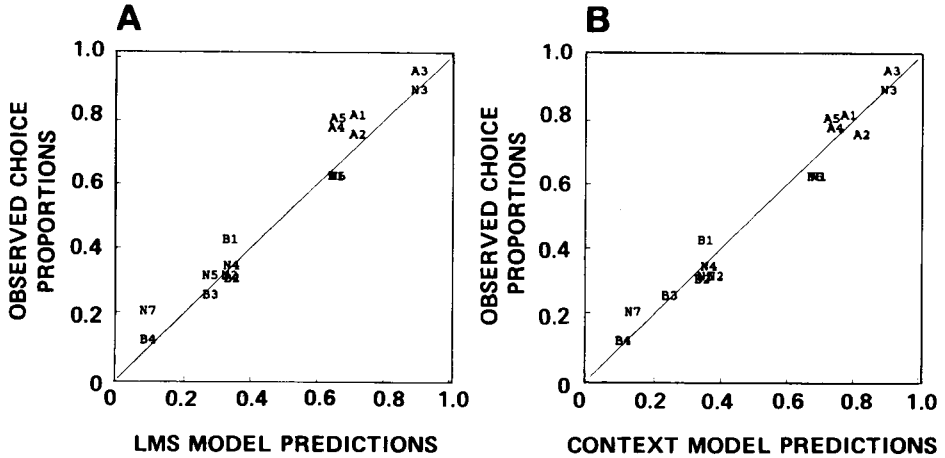

FIG. 8. Choice probabilities from one of the conditions ("last name infinite") reported by Medin et al. (1983) showing comparisons of the predictions (A) the LMS rule with no free parameters and (B) the context model with four similarity parameters estimated from the data. A1 through A5 refer to instances of the A category, and B1 through B4 are instances of the B category. Each instance corresponds to a vector of binary values on four stimulus dimensions. See Table 1 for the identities of the different exemplar types.

weights such that the members of each category are tightly clustered about a central prototype. In this case, the weights discovered by the LMS procedure correspond closely to those found by the Nosofsky optimization procedure. The only difference arises because Nosofsky minimizes percentage errors, whereas the LMS scheme minimizes the square of the difference between the activation sum entered into the logistic and a target of 1 or -1. In this case, the parameter θ corresponds to the parameter D in Nosofsky's optimization. The approximation improves further when the weights of the network model are chosen so as to minimize the sum squared error of the output rather than the activation sums. In any event, the close fit of the LMS rule to Nosofsky's attention optimization model suggests examining other successful applications of Nosofsky's model (1986). We are continuing these theoretical investigations.

Hypothesis-Testing Models

Historically, notions of selective attention in human discrimination learning have had greatest currency in hypothesis-testing models of concept identification. Such models were proposed and investigated by Bruner, Goodnow, and Austin (1956), Restle (1962), Bower and Trabasso (1964), and Trabasso and Bower (1968). The typical concept identification task presents subjects with stimuli such as geometric patterns varying in N binary dimensions (shape, size, color, position, etc.) and asks them to learn a simple classification rule (e.g., "Large figures are As, small figures are Bs"). The basic hypothesis model for such tasks assumes that a subject selects and tries out different hypotheses regarding the correct rule. The hypotheses are coordinated with exclusive attention to one stimulus (e.g., the color or shape dimension); and the hypothesis is selected from a small set of hypotheses (e.g., the $2N$ one-dimension classificatory rules). The model assumes that the subject starts each trial with one hypothesis, uses it to notice the hypothesized feature of the presented stimulus pattern, and classifies the pattern according to the hypothesis. The hypothesis is either maintained, reversed, or rejected depending on the feedback for a given choice.

A lingering difficulty for these simple hypothesis-testing models was their inability to deal adequately with people's learning to identify concepts involving more than simple, one-dimensional rules. In order to come close, the simpler models had to assume that subjects were systematically sampling from the complete "power set" of all 2^{N+1} possible hypotheses involving single stimulus values, doublets, triplets, etc., in various logical combinations (see Hayes-Roth & Hayes-Roth, 1977; Hunt, 1962; Reitman & Bower, 1973; for exceptions, see Bourne, 1970; Hunt, Marin, & Stone, 1962). It was known that learning was more difficult the greater the number of stimulus dimensions (values) that had to be included in the correct rule. The experiment by Shepard, Hovland, and Jenkins (1961), to be described below, provides one demonstration of this fact. Nevertheless, there has been no simple, compelling explanation for that elementary fact within hypothesis-testing theories. This deficiency, along with the upsurge of interest in models for learning ill-defined concepts, caused interest in hypothesis models to wane over the past decades.

Complex concept problems require the subject to base the response on two or more dimensions of stimulus variation. Nosofsky has applied his attention model to the learning of concept problems of differing complexities, estimating how subjects distribute their attention across relevant and irrelevant stimulus dimensions for such complex problems. For example, he fit his model to the results of the experiment of Shepard et al. (1961) to be described below. However, while his model

can *describe* this optimal attention distribution at the asymptote of learning, it does not explain how the subject comes to learn this attention distribution. Also, in order for the model to predict eventual errorless performance, Nosofsky must assume that the stimuli become very dissimilar as learning progresses.

We wanted to see whether our network model can be upgraded to account for the salient findings on problem difficulty. We will begin by fitting the network model to the concept-learning data of Shepard et al., to which we now turn.

Shepard et al. investigated adults' ability to learn to classify eight stimuli comprised of three separable, binary dimensions. Such a stimulus set can be partitioned into two categories of four exemplars each in 70 different ways. But there are really only six distinct types of classifications if we ignore the identity of the three dimensions. For example, a categorization separating the four large from the four small stimuli would be structurally equivalent to one in which the four black and four white stimuli are categorized differently. Figure 9 shows one example of each of the six classification types. In type I classifications, only one dimension is relevant and two are irrelevant; in type II classifications, two dimensions are relevant; in type VI classifications, all three dimensions are equally relevant. Types III, IV, and V are intermediate in complexity between type II and type VI.

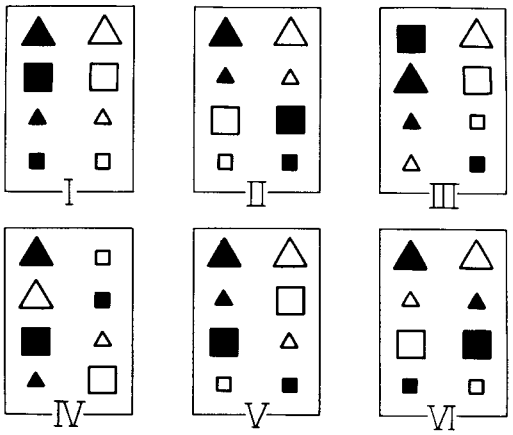

FIG. 9. Examples of the six types of classifications used by Shepard et al. (1961). From "Learning and memorization of classification," by Shepard, et al. (1961). *Psychological Monographs*. Reprinted by permission.

Shepard et al. trained subjects to classify the eight exemplars into their appropriate binary categories using standard procedures: subjects saw the eight stimuli in repeated cycles, one at a time, assigning each to category 1 or 2, and received feedback about the correct answer. The six different problem types of Fig. 9 were compared in speed of learning. A major finding of the experiment was a consistent ordering of difficulty of the classification types. For both trials to learn the classification and total errors made during learning, the order of difficulty of the problem types was (easiest) I < II < III, IV, V < VI (hardest).

Shepard et al. explained their results in terms of supposed attentional effects in the different problems. Using data from subjects' confusion errors and trials to criterion learning, Shepard et al. showed that the different tasks induced behavior which one could interpret as the subjects selectively attending to one or more dimensions as demanded by the task. For example, in type I problems, subjects appeared to be attending predominantly to the one relevant dimension (brightness in Fig. 9); in type II problems, subjects appeared to attend predominantly to the two relevant dimensions (e.g., brightness and shape in II of Fig. 9). Nosofsky (1984) showed that this particular distribution of attention across problem types could be described by assuming that subjects attend selectively to relevant dimensions so as to optimize their performance. However, we will show below that the LMS rule for such tasks also provides a distribution of strengths (attention) which is similar to the optimum characterized by Nosofsky.

On Nonlinear Classifications

While classification types I and IV are

linearly separable tasks, types II, III, V, and VI are not. The simple model in Fig. 1 cannot achieve errorless performance on such problems. Network models have two ways to handle such tasks. First, theorists can postulate intermediate (hidden) units which receive input from each of the stimulus values on the dimensions. Figure 10 shows a possible two-layer network. For example, the presentation of a small white square would cause activity to occur on three input nodes corresponding to small, white, and square and these activations would be sent on to intermediate units which code conjunctions such as small square. If the classification depends on a disjunction of the conjunction of features, such as having black triangle or white square both assigned to the same category (a type II problem), then one expects different hidden units to end up effectively coding each possible conjunction of the relevant single features, with these in turn being strongly connected to the appropriate output (category) nodes. These hidden units act like filters that create internal codes, in that they reduce full patterns to units that fire only for the relevant cues, thus stripping away the irrelevant cues.

Numerous demonstrations have shown that such multilayered networks have great power for learning complex discriminations. In fact, most of the extant network models use two or more layers. However,

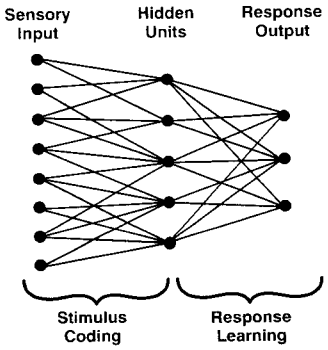

FIG. 10. Diagram of a hypothetical two-layer network which encodes stimuli in the first stage and then selects a response in the second stage.

before we rushed to embrace such a two-layer network approach, we wanted to first see how well the results could be fit with a patched up version of a one-layer network which we call the configural-cue model.

The Configural-Cue Model

The basic idea here is to expand the definition of the input stimulus to include sensory nodes that correspond to each possible single feature, conjunction of features, triplets of features and so on, all of these in a one-layer network. For three binary dimensions, there are 26 such input codes. Let us assume that the appropriate nodes are turned on at the first (sensory) layer when the corresponding stimulus pattern is presented. Presentation of an n-dimensional stimulus pattern would then correspond to presentation of the complete power set of all possible subsets of that pattern, each being associated to whatever category is correct on that trial. Thus, 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, e.g., Fig. 11). This power set coding of stimuli at the input layer essentially takes hidden units for all possible conjunctive combinations of cues and brings them out to the sensory input layer. The method appears to be quite wasteful of coding units, after all, codes are created for every possible feature configuration, whether or not the code is ever needed. Whereas multilayered networks economize by having fewer sensory units and tuning some nonspecific hidden units to code needed conjunctions, the configural cue model economizes by simplifying the computation of changes in the weights during learning. The unpalatable aspect of the configural cue model is the exponential growth of the number of possible configurations as the number of stimulus dimensions becomes larger. In practice, we have used the model only with the learning of

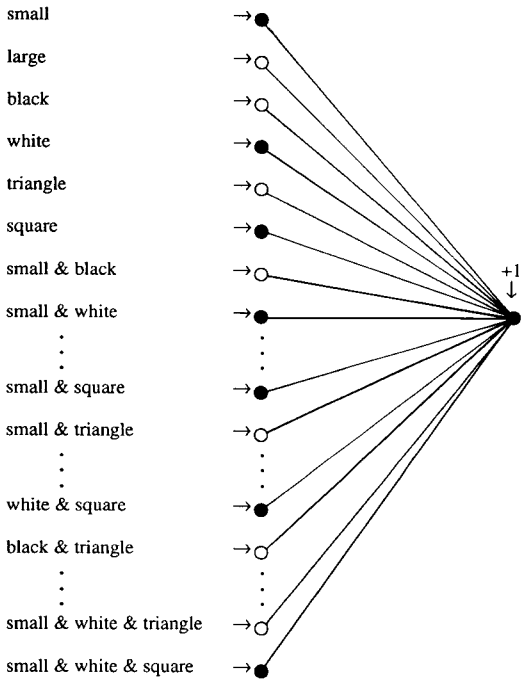

FIG. 11. A one-layer configural cue network for modeling learning of the Shepard et al. (1961) tasks. The solid input nodes correspond to the seven nodes activated by presentation of a small white square: small, white, square, small white, small square, white square, and small white square. Note that only a subset of all 26 input nodes are shown.

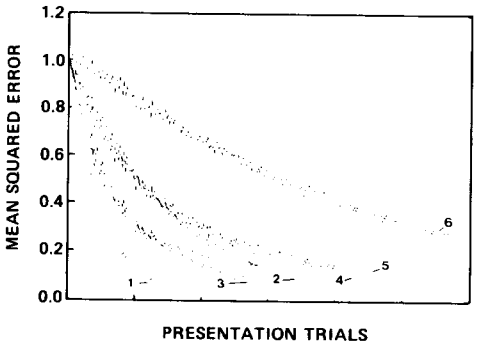

FIG. 12. Mean squared error plotted over trials from simulations of learning the six classification types of Shepard et al. (1961) using a one-layer configural cue network. Each curve is based on 300 simulated subjects in that condition using a β of .02.

stimulus patterns varying in less than four dimensions.

To check the predictions of the configural cue model, we simulated the Shepard et al. study using the LMS rule for modifying weights. For each classification type, the network was presented many times with randomly chosen exemplars from the sets of eight in Fig. 9, each paired with the correct classification. To economize on the millions of computations needed for 300 stat-subjects per six conditions, we used as our measure of learning the mean squared error (averaged across 300 Monte Carlo simulations) throughout training. Figure 12 graphs the mean squared error over training trials for each condition; this measure is closely related to the percentage of incorrect classifications.

Examining the simulated curves in Fig. 12, one is led to several conclusions. First, by any criterion, problem I is clearly the easiest and problem VI the hardest. Second, the relationship between problems II, III, IV, and V depends on which stage of practice is selected for comparison. During early trials, problems III and IV are easier than II and V; later, problems II and IV interchange their order, so that the overall ordering late in practice is I < III < II < IV < V < VI. The initial ordering of the problems is understandable in terms of the number of partially valid cues that enable above chance performance. Examining the partitions in Fig. 9, the numbers of partially valid cues are zero for problem type II, one (color) for type V, two (color and shape) for type III, and three (color, shape, and size) for type IV. Early in training the network will acquire large weights to these partially valid single cues, enabling a rapid early decline in errors; however, this low error rate will delay complete learning, especially in problems types IV and V.

The crossovers in the learning curves predicted in Fig. 12 cannot be tested on the data of Shepard et al., since learning curves were not reported. If we were to measure learning rate by the trials before the average error rate attained some small percentage (say, 10%), then our simulations

imply about the same ordering of the six problems as observed by Shepard et al., except that types II and III are reversed. We are unsure how much weight to place on this discrepancy; the original data were obtained in a complex within-subjects experiment in which only a few subjects learned six replications of all six problem types in Latin-square order and substantial practice effects occurred. Thus, a replication of the basic experiment would tell us whether this misordering is reliable.

The asymptotic weights to which the network converges for the different problems reflect the S–R contingencies for those problems. Thus, for type I, the model converges to large weights for the single relevant cues such as black and white but small weights everywhere else. For the type II problem, it converges to large weights on the relevant conjunctions, such as *black triangle* and *white square,* but it has small weights elsewhere. For type VI, it converges to large weights on the triplets and small weights elsewhere.

Since the triplet input nodes are always perfectly correlated with the correct category, one might wonder why these triplets do not pick up all the conditioning and swamp out any of the singlet or doublet cues, even for the simple problems. This scenario is prevented by the competitive nature of the LMS learning rule. Consider the case of a type I problem where the network sees a small white square; this turns on the input units for the single cues and the double cues, as well as the triplet cue (see Fig. 11). If white is perfectly relevant, then it alone has occurred always paired with category 2 four times as frequently as the small white square unit; hence, the association or weight of the single-feature node will be much stronger. Therefore, by the LMS rule, the weight for the *small white square* unit will not get much of a boost in its association to category 2 because that category is already being strongly predicted by the *white* cue alone.

Similarly, in problems of type II, the relevant doublet cues beat out the triplet cues because they occur twice as often as the triplets, and so their conditioning will block that of the triplets. Of course, the doublets in type II problems beat out the single features because the single features alone are uncorrelated with the correct classification in this case.

The general principle is that the LMS rule assures that a more frequent valid cue will block or beat down the conditioning of another valid cue that appears less often. It is this subtle aspect of the LMS algorithm, reminiscent of the overshadowing and blocking effects in classical conditioning, that allows the configural cue model to display a compelling feature of human learning, namely, subjects' apparent drift toward "increasing complexity of hypotheses testing" in which simple hypotheses or cues *appear* to be tested before more complex hypotheses.

Earlier we noted that an inelegancy of this configural cue model is that the number of distinct subpatterns to be tracked increases dramatically with the number of stimulus dimensions. But in further explorations, we have discovered that the model can do remarkably well in fitting data if we expand the single-element coding to allow merely doublet coding of simple conjunctions. For example, a model with only singlet and doublet coding does a good job accounting for the results which Medin and Schaffer (1978) used to reject single-cue (independent) models. Of course, problems of type VI which require subjects to take account of triplets of features would create difficulties for the doublet model.

Two-Layered Models

While the configural cue model provides a fairly good account with *no* parameters of the difficulty ordering of problems of Shepard et al., it clearly encounters several conceptual difficulties in dealing with other discrimination learning data. Interestingly,

its difficulties are exactly those raised historically by two-process discrimination learning theorists (e.g., Kendler & Kendler, 1962; Lawrence, 1949, 1950; Sutherland & Mackintosh, 1964; Zeaman & House, 1963) in arguing against one-process discrimination learning theories such as Spence's (1936) or Estes and Burke's (1953). The two-process theorists argued that discrimination learning involved the subjects (1) learning to code the stimuli according to the relevant dimension, and (2) learning which overt responses to make to the stimuli-as-coded. The configural cue model conflates and confuses these two processes, and this creates specific difficulties for it.

The basic difficulty arises from results showing that subjects (animals or humans) can learn about the relevance or irrelevance of stimuli somewhat independently of learning what responses to make to these stimuli. For example, Lawrence's (1949, 1950) classic experiments on "acquired distinctiveness of cues" demonstrated that rats could learn to attend to, for example, brightness as a relevant dimension and ignore, for example, shape as an irrelevant dimension in one problem, and later transfer that learning to a new problem involving similar cues but completely different responses. Zeaman and House's (1963) demonstration that intradimensional shifts are easier than extradimensional shifts, or Kendler and Kendler's (1962) demonstration that reversal shifts are usually easier than nonreversal shifts, make the same point theoretically.

We would propose that the predictions of such two-process models might be approximated by the two-layer network models. Such a network is depicted in Fig. 10. We can think of the intermediate layer of hidden units as representing different possible stimulus codes or filters for the first layer, which only register the full sensory pattern. Thus, the links from the sensory layer to the hidden units reflect the first process, the learning of stimulus codes, in the two-process theories. The links from the hidden layer to the output layer would then reflect the second process, the associations between the stimuli as coded and the overt responses.

Such a model should have the potential to deal with the results mentioned above. For example, the model should learn in an initial problem to code relevant cues and ignore irrelevant cues; and it can then lock in (clamp) these intermediate codes while learning different responses to these stimuli in a second problem. This then could explain the results on reversal vs nonreversal shift, and on acquired distinctiveness of cues, though not on intradimensional shifts. Two-layer models seem also to show insightful learning curves for difficult concept problems; the lengthy presolution level of chance responding corresponds to the model learning slowly the proper coding for the hidden units. Furthermore, the model could learn to cluster together sets of stimulus patterns on the basis of their internally correlated features, even in the absence of feedback information about a classification (see, e.g., Hanson & Kegl, 1987). Psychologists call this perceptual learning, predifferentiation, or prefamiliarization with the stimulus set. Prefamiliarization usually facilitates later learning of discriminative responses to the stimuli (see review by Gibson, 1969).

A problem with the general two-layer model is that it is almost too flexible and unconstrained. A number of distinct models, such as the two-process model noted above, arise as special cases of the general model. As a second example, one can cast Medin's context model in this two-layer format, where each distinct training pattern turns on its unique hidden unit, which is then connected to the correct response. The weights then reflect how much the different values of a given stimulus dimension behave similarly in activating hidden units corresponding to dif-

ferent training patterns. The fact that the same framework can be used for stating such disparate models as the two-process attention model and Medin's context model tells us that the general framework does not materially constrain which of several instantiations can be dervied within the framework. In any event, we are currently exploring by simulation the behavior of simple two-layered learning networks and trying to relate the simulations to results on discrimination learning. Those investigations are at too early a stage to report.

Discussion of Attentional and Hypothesis-Testing Model Explorations

To review the ground covered in this section, we began by questioning whether the network model using the LMS learning rule could reproduce some apparent attention-like phenomena from discrimination/classification learning. First, we found that even the one-layer model could do a reasonable job of mimicking the weights (parameters) of Nosofsky's optimal attention model for some experiments. But our weights, which the LMS rule calculates a priori from the stimulus structure of the classification, are interpreted as stimulus–response associations rather than as attentional saliencies. Moreover, for a class of simple problems, the weights obtained by the LMS rule are very close to those of Nosofsky's optimization model.

Second, we showed that a one-layered network model with an expanded set of sensory inputs could approximate the order of learning difficulty for the six logical types of problems in the Shepard et al. (1961) experiment. However, we noted that this configural cue model encounters problems in explaining classical results revealing the learning of stimulus codes which extract relevant stimulus dimensions from the set of exemplars. We then speculated that two-layer networks, with hidden units that reductively code the relevant features or featural combinations, might have the explanatory power of the earlier two-process theories (e.g., Sutherland & Mackintosh, 1971; Zeaman & House, 1963). Moreover, the two-layer network models probably have even greater discriminative powers because they can develop intermediate codes for whatever featural configurations are significant (correlated) for the induction task at hand, whereas the earlier two-process theories really only dealt with selective coding of single dimensions (e.g., "look at the *shape* of the geometric figure") and were silent about how more complex stimulus codes arose from the structure of the experienced stimulus set.

Concluding Remarks

This paper recounts our explorations of the power of the LMS rule in the context of a one-layer network. The early work showed the virtues of the one-layer network, especially the LMS learning rule. We are also impressed with its explanation of some attentional phenomena in discrimination learning, especially when it was formulated as the configural cue model. Eventually, however, arguments were marshalled against both one-layer models and in favor of at least a two-layer model.

The reader may wonder why we even bothered investigating the one-layer model when logical arguments against it have been known for many years (e.g., Minsky & Papert in 1969 reviewed its deficiencies). We believe that the one-layer model should remain a viable candidate in the tool kit of the theoretical psychologist for several reasons. A first reason is the utter simplicity of the one-layer model (akin to the simple all-or-none model of an earlier era, see Bower, 1961; Estes, 1960). Psychologists are attracted to simple models that provide "back of the envelope" quick answers, that are frequently close to the facts of the case. Simple models are easy to remember, make strong predictions (often parameter-free), and are useful for ex-

ploring how much of the variance in data can be explained by elementary learning processes and highlighting discrepancies which point the direction toward more sophisticated models. Simple models also exemplify the canon of parsimony so dear to the hearts of experimental psychologists, viz., do not propose a more complex hypothesis than is required to explain the data at hand.

A second reason is that by connecting the Rescorla–Wagner model of conditioning to phenomena of human learning, the one-layer model suggests the possibility that the LMS rule may form an algorithmic building block for human associative learning. As noted earlier, the assumption of phyletic continuity underlay much of the early interest in animal conditioning. A renewed attempt at deriving processes of complex learning from configurations of the elementary associative processes observed in lower animals seems especially timely given recent advances in identifying and modeling the neural substrates of simple forms of associative learning (e.g., Gluck & Thompson, 1987; Thompson, 1986).

A third reason for using the simple model is that under certain conditions the more complex, multilayer networks end up making nearly the same predictions as does the one-layer network, although they arrive there by a more complex route. If each hidden unit produces an output that is a linear function of its input, then the whole cascade is linear, and so it can be mimicked by a one-layer network with suitable weights. We have found in simulations that even with nonlinear hidden units (e.g., whose output is a logistic function of their input activations), if all are equally connected to the input units and if all inputs have *independent* correlations with the desired outputs, then a one-layer network often provides a close approximation to the predictions of a two-layer network.

Obviously, we are in the beginning stage of our investigations of the explanatory and predictive power of the LMS learning rule within one-layer and two-layer associative networks. We are especially pleased that these models link up naturally with two venerable traditions in learning theory, namely, studies of elementary association formation in animal conditioning studies, and potentially with the two-process theories of discrimination learning and hypothesis testing. It is gratifying to see that classic results and arguments from old debates are just as forceful and pointed today in a modern theoretical framework as they were decades ago when first published. It provides workers and students with a sense of continuity of concern with certain fundamental problems in the field as well as a belief in cumulative development and convergence of different theoretical trends. The authors are pleased to see that a form of associationism is returning to being a theoretical contender, if not yet a popular one, after having been earlier consigned to the flames by sweeping polemics.

We are pleased that our minimal one-layer model has taken us so far in explaining interesting behavioral results and can guide new experiments whose results confirm this model and present challenges to popular competing theories. Our guiding principles have been simplicity and economy. We know the simple model may be approximately correct only in restricted experimental circumstances, but we feel it is helpful in exploring this fascinating terrain. Its successes and failures inform us and materially constrain the class of more sophisticated models that will be needed to explain an increasingly complex pattern of results.

REFERENCES

ATKINSON, R. C., & ESTES, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.

BAKER, A. G., & MACKINTOSH, N. J. (1977). Excitatory and inhibitory conditioning following uncor-

related presentations of CS and UCS. *Animal Learning and Behavior*, **5**, 315–319.

BOURNE, L. (1970). Knowing and using concepts. *Psychological Review*, **77**, 546–556.

BOWER, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika*, **26**, 255–280.

BOWER, G. H., & HILGARD, E. R. (1981). *Theories of learning*. New Jersey: Prentice–Hall.

BOWER, G. H., & TRABASSO, T. (1964). Concept identification. In R. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford, CA: Stanford Univ. Press.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. (1956). *A study of thinking*. New York: Wiley.

CASTELLAN, N. J. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, D. P. Pisoni, & G. R. Potts (Eds.), *Cognitive Theory* (Vol. 2). Hillsdale, NJ: Erlbaum.

COTTRELL, G. (1985). *A connectionist approach to word sense disambiguation* (Report #154). Rochester, NY: University of Rochester, Computer Science Department.

DELL, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283–321.

DICKINSON, A., HALL, G., & MACKINTOSH, N. J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, **2**, 213–222.

ELMAN, J. L., & MCCLELLAND, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, **27**, 143–165.

ESTES, W. K. (1960). Learning theory and the new mental chemistry. *Psychological Review*, **67**, 207–223.

ESTES, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, **67**, 81–102.

ESTES, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500–549.

ESTES, W. K., & BURKE, C. J. (1953). A theory of stimulus variability. *Psychological Review*, **60**, 276–286.

FRANKS, J. J., & BRANSFORD, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, **90**, 65–74.

FRIED, L. S., & HOLYOAK, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 234–257.

GIBSON, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton–Century–Crofts.

GLUCK, M. A., & BOWER, G. H. (in press). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*.

GLUCK, M. A., & THOMPSON, R. F. (1987). Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review*, **94**, 176–191.

HANSON, S. J., & KEGL, J. (1987). Language grammar from exposure to natural language. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society, Seattle, WA*.

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

HINTON, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Amherst, MA*.

HINTZMANN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **934**, 411–428.

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

HULL, C. L. (1943). *Principles of behavior*. New York: Appleton–Century–Crofts.

HUNT, E. B. (1962). *Concept learning*. New York: Wiley.

HUNT, E. B., MARIN, J., & STONE, T. J. (1962). *Experiments in induction*. New York: Academic Press.

KAHNEMAN, D., & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–251.

KAMIN, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton–Century–Crofts.

KENDLER, H. H., & KENDLER, T. S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review*, **69**, 1–16.

KOHONEN, T. (1977). *Associative memory: A system-theoretic approach*. New York: Springer-Verlag.

KREMER, E. F. (1978). The Rescorla–Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, **4**, 22–36.

LAWRENCE, D. H. (1949). Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, **39**, 770–784.

LAWRENCE, D. H. (1950). Acquired distinctiveness of

cues: II. Selective associations in a constant stimulus situation. *Journal of Experimental Psychology*, **40**, 175–185.

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

MACMILLAN, J. (1987). *The role of frequency memory in category judgments*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

MCCLELLAND, J. L., & ELMAN, J. L. (1986). Interactive processes in speech perception: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

MCCLELLAND, J. L., & RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375–407.

MCCLELLAND, J. L., & RUMELHART, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

MCCULLOCH, W. S., & PITTS, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133.

MEDIN, D. L., ALTOM, M. W., EDELSON, S. M., & FREKO, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **8**, 37–50.

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

MEDIN, D. L., & EDELSON, S. M. (in press). Problem structure and the use of base rate information from experience. *Journal of Experimental Psychology: General*.

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

MEDIN, D. L., & SCHWANENFLUGEL, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 355–368.

MEDIN, D., & SMITH, E. (1984). Concepts and concept formation. *Annual Review of Psychology*, **35**, 112–138.

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

NEUMANN, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory and Cognition*, **2**, 241–248.

NOSOFSKY, R. (1988). Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 54–65.

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

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

PARKER, D. (1985). *Learning logic* (Report #47). Cambridge, MA: MIT, Center for Computational Research in Economics and Management Science.

PARKER, D. (1986). A comparison of algorithms for neuron-like cells. In *Proceedings of the Neural Networks for Computing Conference, Snowbird, Utah*.

PEARCE, J. M., & HALL, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned and unconditioned stimuli. *Psychological Review*, **87**, 532–552.

POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353–363.

RASHEVSKY, N. (1937). Mathematical biophysics of conditioning. *Psychometrika*, **2**, 199–209.

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

REITMAN, J. S., & BOWER, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology*, **4**, 194–206.

RESCORLA, R. A. (1971). Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learning and Motivation*, **2**, 113–123.

RESCORLA, R. A., & HOLLAND, P. C. (1982). Behavioral studies of associative learning in animals. *Annual Review of Psychology*, **33**, 265–308.

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

RESTLE, F. (1962). The selection of strategies in cue learning. *Psychological Review*, **69**, 11–19.

ROITBLATT, H. L. (1987). *Introduction to comparative cognition*. New York: Freeman.

ROSENBLATT, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of the brain mechanisms*. Washington, DC: Spartan.

RUMELHART, D. E., & MCCLELLAND, J. L. (1986a).

On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

RUMELHART, D. E., & MCCLELLAND, J. L. (1986b). *Parallel Distributed Processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. Cambridge, MA: Bradford Books/MIT Press.

SEJNOWSKI, T. J., & ROSENBERG, C. R. (1986). *NETtalk: A parallel network that learns to read aloud* (Technical Report JHU/EECS-86/01). Baltimore, MD: Johns Hopkins University.

SHEPARD, R. N., HOVLAND, C. I., & JENKINS, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**, 1–42.

SLOVIC, P., & LICHTENSTEIN, S. C. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, **6**, 649–744.

SPENCE, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review*, **43**, 427–449.

STONE, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. Cambridge, MA: Bradford Books/MIT Press.

SUTHERLAND, N. S., & MACKINTOSH, N. J. (1964). Discrimination learning and the non-additivity of cues. *Nature (London)*, **201**, 528–530.

SUTHERLAND, N. S., & MACKINTOSH, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

SUTTON, R. S., & BARTO, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, **88**, 135–170.

THOMPSON, R. F. (1986). The neurobiology of learning and memory. *Science*, **233**, 941–947.

THURSTONE, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273–286.

TRABASSO, T., & BOWER, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.

TVERSKY, A. (1977). Features of similarity. *Psychological Review*, **84**, 327–352.

TVERSKY, A., & KAHNEMAN, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases*. Cambridge: Cambridge Univ. Press.

WAGNER, A. R. (1971). Elementary association. In H. H. Kendler & J. T. Spence (Eds.), *Essays in neobehaviorism: A memorial volume to Kenneth W. Spence*. New York: Appleton–Century–Crofts.

WAGNER, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. Spear & G. Miller (Eds.), *Information processing in animals: Memory mechanisms*. Hillsdale, NJ: Erlbaum.

WAGNER, A. R., & RESCORLA, R. A. (1972). Inhibition in Pavlovian conditioning: Applications of a theory. In R. A. Boakes & S. Halliday (Eds.), *Inhibition and learning* (pp. 301–336). New York: Academic Press.

WALTZ, D. L., & POLLACK, J. B. (1985). Massively parallel processing. *Cognitive Science*, **9**, 51–74.

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

ZEAMAN, D., & HOUSE, B. J. (1963). The role of attention in retardate discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency*. New York: McGraw–Hill.

ZIMMER-HART, C. L., & RESCORLA, R. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, **86**, 837–845.

(Received June 2, 1987)
(Revision received November 25, 1987)