Journal of Experimental Psychology: General
1988, Vol. 117, No. 3, 227-247

0096-3445/88/$00.75

From Conditioning to Category Learning: An Adaptive Network Model

Mark A. Gluck and Gordon H. Bower
Stanford University

We used adaptive network theory to extend the Rescorla–Wagner (1972) least mean squares (LMS) model of associative learning to phenomena of human learning and judgment. In three experiments subjects learned to categorize hypothetical patients with particular symptom patterns as having certain diseases. When one disease is far more likely than another, the model predicts that subjects will substantially overestimate the diagnosticity of the more valid symptom for the rare disease. The results of Experiments 1 and 2 provide clear support for this prediction in contradistinction to predictions from probability matching, exemplar retrieval, or simple prototype learning models. Experiment 3 contrasted the adaptive network model with one predicting pattern-probability matching when patients always had four symptoms (chosen from four opponent pairs) rather than the presence or absence of each of four symptoms, as in Experiment 1. The results again support the Rescorla–Wagner LMS learning rule as embedded within an adaptive network model.

To what extent do the processes of human learning emerge from complex configurations and elaborations of the elementary learning processes observed in animals? Research in the two areas of human and infrahuman learning shares a long history that has focused on elementary associative learning (Ebbinghaus, 1885; Pavlov, 1927). About 20 years ago, however, animal and human learning research became divorced from each other. Animal research continued to be primarily concerned with elementary associative processes (Mackintosh, 1983; Mackintosh & Honig, 1969; Rescorla & Holland, 1982), whereas human learning (or memory) tended to be characterized in terms of information processing and rule-based symbol manipulation, an approach borrowed from artificial intelligence.

In spite of this divergence, a number of recent empirical findings suggest that there may be some common aspects to human and animal learning (e.g., Alloy & Tabachnik, 1984; Dickinson & Shanks, 1985; Estes, 1985; Medin & Dewey, 1984). Furthermore, interest in relating human cognition to configurations of elementary associative connections has recently been revived by the development of adaptive networks as models of cognitive processes. Such models—also known as parallel distributed processing or connectionist networks—are being developed to simulate diverse cognitive behaviors such as learning, pattern recognition, speech recognition and production, motor control, and so on (e.g., Hinton & Anderson, 1981; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). This theoretical movement represents a return to a traditional goal of psychology, which was to view complex human abilities as emerging from configurations of elementary associative processes that could be studied in simple organisms.

Given the voluminous studies of learning in animals alongside current attempts to model cognition with elementary associative processes, it seems particularly timely to search for and exploit any correspondence that may exist between animal and human associative learning.

Associative Learning in Animals

The most basic and well-studied form of learning is classical Pavlovian conditioning. In classical conditioning a previously neutral stimulus—the *conditioned stimulus* (CS), such as a bell—comes to be associated with a biologically significant stimulus—the *unconditioned stimulus* (US), such as food or an electric shock. Early learning theories assumed that the simple temporal contiguity or joint occurrence of a CS and US was sufficient for associative learning (e.g., Hull, 1943; Spence, 1956). Later experiments made clear, however, that simple contiguity was not sufficient. The ability of a CS to become conditioned to a US depended on its impairing reliable and nonredundant information about the occurrence of the US (Kamin, 1969; Rescorla, 1968; Wagner, 1969).

A critical observation suggesting this principle was an experiment on blocking of conditioning (Kamin, 1969). In Kamin's experiment, a light, the CS, was conditioned to predict a shock, the US. Then a compound stimulus consisting

This work was partially supported by research grants to Gordon H. Bower from the National Institute of Mental Health (MH-13950), from the National Science Foundation (BNS-86-18049), and from the Sloan Foundation.

For sharing with us his insights into the formal properties of the models presented in this article and for his invaluable assistance in preparing the appendixes, we are especially indebted to David Parker. For their helpful comments on earlier versions, we thank Andy Barto, Nelson Donegan, W. K. Estes, Daniel Kahneman, Jay McClelland, Mike McClosky, Robert Nosofsky, Misha Pavel, Paul Rosenbloom, Roger Shepard, Richard Sutton, Richard Thompson, and Amos Tversky. For their assistance with this research, we are grateful to Thom Deaton, Andrea Gallagher, Carrie Henderson, Anthony Henin, Van Henkle, Robert Kylberg, Gus Larsson, Susanna Lee, Naomi Schechter, and Jamie Uyehara.

Correspondence concerning this article should be addressed to Mark A. Gluck, Stanford University, Department of Psychology, Jordan Hall, Building 420, Stanford, California 94305-2130.

of a light and a tone was paired with the shock. Surprisingly, learning the tone → shock association hardly occurred at all compared with control subjects who had received no pretraining to the light. This result, similar to Pavlov's (1927) work on the overshadowing of one by another, is called *blocking* because prior training of the light → shock association blocks learning of the tone → shock association during the second, (light + tone) → shock stage of training.

The blocking effect suggested that the effectiveness of a US for producing associative learning depends on the relationship between the CS and the expected outcome (Kamin, 1969; Rescorla, 1968; Wagner, 1969). Rescorla and Wagner provided a precise formulation of this proposal (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). Their formation assumed that the association that accrues between a stimulus and its outcome on a trial is proportional to the degree to which the outcome is unexpected (or unpredicted) given all the stimulus elements that are present on that trial. To formulate the relationship, let V_i denote the strength of association between stimulus element CS_i and the US. If CS_i is followed by a reinforcing unconditioned stimulus, then the change in the associative strength between CS_i and the US, ΔV_i, can be described by Equation 1:

$$\Delta V_i = \alpha_i \beta_1 (\lambda - \sum_{k \in S} V_k), \tag{1}$$

where α_i reflects the intensity or salience of CS_i, β_1 reflects the rate of learning on US trials, λ is the maximum possible level of associative strength conditionable with that US intensity, and $\sum_{k \in S} V_k$ is the sum of the associative strengths between all the CS stimulus elements occurring on that trial and the US. If CS_i is presented on a trial and not followed by the US, then the association between CS_i and the US decreases analogously, namely,

$$\Delta V_i = \alpha_i \beta_2 (0 - \sum_{k \in S} V_k), \tag{2}$$

where 0 is the level of associative strength supported by nonpresentation of the US and β_2 reflects the rate of change of the association due to nonreinforcement. Generally, β_1 is assumed to be larger than β_2, but this is not critical for most predictions (see Rescorla & Wagner, 1972).

The Rescorla–Wagner (1972) model is the most widely accepted description of associative changes during classical conditioning. The wealth of confirmed implications arising from this deceptively simple model has been substantial.[1] This model accounts for the blocking effects as follows: When in Phase 1, CS_i has been initially conditioned to the US, V_1 approaches λ. If the initial associative strength, V_2, of the novel stimulus is assumed to be zero, then the compound stimulus strength, $V_1 + V_2$, will equal λ. By Equation 1, when the compound is paired with the US, the incremental learning accruing to the novel stimulus, ΔV_2, is thus predicted to be zero—as observed. By a similar logic, the model also predicts the results of experiments in which single CS_1 – US trials are interspersed among compound trials in which CS_1 and CS_2 are presented simultaneously and paired with the US. In such conditions, the novel element on compound trials, CS_2, acquires relatively little associative strength. This is because the single CS_1 – US trials strengthen V_1; and this, by Equation 1, reduces the increment available to V_2 on compound, (CS_1 + CS_2) – US, trials.

The model also provides an explanation for a related finding by Rescorla (1968), who demonstrated that the conditioning of a CS to a US depends on the probability of the US occurring in the presence of the CS relative to the probability of the US occurring in the experimental situation but without the CS. He found that conditioning proceeds to a level proportional to the contingency (or correlation) between the tone and the shock and is not solely related to the conditional probability of the US given the CS. These results are consistent with the Rescorla–Wagner (1972) model if CS_1 is identified as the background stimulus of the conditioning box and CS_2 is identified with the added tone, thus making up the conceptual compound $CS_1 + CS_2$. Increasing the US rate to CS_1 alone (i.e., unpredicted USs) will—by arguments similar to those above—increase V_1, thereby blocking the development of associative strength to the CS_2 on (CS_1 + CS_2) – US trials. Thus the level of conditioning to the tone (CS_2) will vary with the probability of the US in the presence of the tone compared with the US probability in the absence of the tone, as Rescorla observed.

Despite the importance of the blocking experiment for theories of associative learning, only a few investigators have carried out bridging experiments from animal to human learning. Rudy (1974) noted a parallel between human paired-associate learning and animal associative learning and pointed to a form of blocking in human learning. Specifically, when a redundantly relevant cue is compounded with stimuli that are already sufficient to evoke the associated response, the added cues are not likely to become associated with the response (Trabasso & Bower, 1968). Dickinson and Shanks (1985) also demonstrated analogues of several conditioning phenomena in human learning. They showed that people's judgments of the correlation of two events are influenced by the conditional status of other events that are present, in a manner reminiscent of blocking and overshadowing phenomena in animal conditioning. Neeley (1982) showed a similar effect in a study in which subjects learned the attractiveness of different political candidates who ran repeated campaigns against each other. Schank (1982) similarly postulated "expectation failure" as the driving force behind conceptual learning; the EPAM model long ago used a similar rule (Feigenbaum, 1959; Feigenbaum & Simon, 1961).

Despite these tantalizing hints that there may be common aspects to human and animal learning, there have been sur-

[1] Despite the many successes of the Rescorla–Wagner (1972) 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).

prisingly few attempts to draw a more rigorous connection between the models of animal learning and human learning. No studies have attempted directly to evaluate whether the Rescorla–Wagner (1972) rule is an appropriate characterization of an algorithm underlying human associative learning.

Adaptive Network Models of Cognition

Recent years have witnessed an increased interest, across the disciplines of cognitive psychology, computer science, and neurobiology, in understanding the information-processing capabilities of complex networks of massively interconnected, neuronlike computing elements. Among theorists studying these network models, the following people are notable for demonstrating the computational power and psychological verisimilitude of these adaptive networks: Ackley, Hinton, and Sejnowski (1985), Hinton and Anderson (1981), McClelland and Rumelhart (1986), and Rumelhart and McClelland (1986).

There are several classes of adaptive networks and modes of processing within these networks. The networks most similar to the models used to describe animal learning are those consisting of processing units connected by weighted unidirectional links (see, however, Ackley et al., 1985, for an alternative class of models). These networks are typically divided into a layer of sensory units; a layer of response units; and zero, one, or more layers of intermediate, association units. The state of each processing unit, at each moment in time, is described by its activation, which is determined by the sum of the weighted inputs to that unit from all its incoming connections. Presentation of a stimulus pattern to the system corresponds to activating a set of sensory units. These units pass their weighted activation along their connections either directly to the output units or to intermediate units that relay them onward, eventually terminating on output units. The activation pattern over the layer of output units corresponds to some particular response of the system to that input. After receiving feedback regarding the desired output pattern for each input pattern, the system adjusts the weights on the connections to have that input produce an output closer to the one desired. By repeatedly cycling through a set of desired input–output pairings, the system "learns" just those weights that will achieve the closest match (of which it is capable) to the input–output pairings. These weights correspond to strengths of associations in classical learning theory, and the algorithm for changing the weights in the light of feedback corresponds to "learning rules" in traditional theories.

We will formulate these ideas more precisely for a one-layer network in which inputs are connected directly to output units. A network is considered to have learned to associate k pairs of patterns, $\{I_1, O_1\} \ldots \{I_k, O_k\}$, if the presentation of $I\alpha$ to the input nodes as a vector of activation produces $O\alpha$ as a pattern of activation on the output nodes. If there are n input nodes and the activation in input node i is α_i, then the activity in output node j, o_j, is determined according to the following rule for the spread of activation from input nodes to output nodes:

$$O_j = \sum_{i=1}^{n} w_{ij} a_i, \tag{3}$$

where the sum is over the n input nodes.

For such a system to be adaptive, the weights, w_{ij}, must be adjusted to map as closely as feasible the k stimuli, I_α, into the corresponding k responses, O_α. A common measure of the accuracy of performance of the network is the expected squared difference between the actual and desired activations at the output nodes. In equation form, this expected squared error, $E[e]$, is equal to

$$E[e] = E\left[\frac{1}{m} \sum_{j=1}^{m} (O_j - d_j)^2\right], \tag{4}$$

where the expectation is taken across the 250 training trials, the summation is over the m output nodes, o_j is the actual activation on output node j, and d_j is the desired output for output node j given the input pattern. Other measures of adaptive accuracy are possible (e.g., minimizing the average percentage errors of the outputs or the expected cost of the errors). In what follows, however, we test the general idea that people learn in such manner as to minimize $E[e]$. Such weights are termed a *least mean squares* (LMS) solution to the problem of associating the k input patterns with their outputs.

There are several "error correcting" learning rules for adjusting the weights so that they converge to an LMS solution. A rule of this kind that has gained considerable popularity among network theorists in recent years is the LMS rule, a variant of the perceptron convergence algorithm (Rosenblatt, 1961), which was first proposed as a learning rule for adaptive networks by Widrow and Hoff (1960). This rule has been called the LMS rule (also the delta rule) because it was the first one discovered that leads to the LMS solution. The LMS rule says that the changes in weights, Δw_{ij}, from input node i to output node j is given by Equation 4:

$$\Delta w_{ij} = \beta(d_j - o_j)a_i, \tag{5}$$

where a_i is the activation on input node i and d_j is a special "teaching" input signal to output node j indicating what the activation of that node should be to obtain the correct response. A number of useful theorems have been proved about the LMS rule (Kohonen, 1977; Parker, 1985, 1986; Stone, 1986). The LMS rule provides a set of linear equations that, when iterated over trials, will converge to weights that will perfectly discriminate among the input patterns (if such weights exist). Otherwise, the algorithm will converge to weights that minimize LMS error between the resulting and desired output patterns (Kohonen, 1977).

One may note the close correspondence among the network activation model, the linear discriminant function approach to classification, and the standard linear-regression model for predicting a criterion variable or category (y) from a set of independent variables, x_i (see Stone, 1986). The regression equation

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \tag{6}$$

has the same form as the summed activation equation with the regression coefficients (b_is) playing the same role as the weights or association strengths (w_is). And like w_i, each b_i reflects the correlation between a predictor variable, x_i, and the criterion, after correcting for intercorrelations among the predictor variables, as does the LMS rule. In particular, the least square estimators of the b_i are equivalent to the asymptotic weights obtained iteratively by changing the w_is trial by trial according to Equation 5 (see, e.g., Stone, 1986). This correspondence implies that a linear-regression model would show many of the phenomena captured by the asymptotic behavior of an LMS network model when applied to a fixed environment (see, e.g., Johnson & Wichern, 1982). There is also an interesting correspondence of the network model to logistic regression and to a Bayesian inference model (see Slovic & Lichtenstein, 1971). We will not pursue these correspondences further here, although they are fascinating topics for exploration.

As noted already, the LMS rule is limited to discriminating linearly separable patterns on the basis of the idea that similar input patterns are mapped to similar output patterns. In cases in which similar input patterns are not mapped to similar output patterns, a layer of additional, intermediate nodes ("hidden units") may be required between the input and output nodes to solve the discrimination (e.g., see Rumelhart, Hinton, & Williams, 1986). Numerous demonstrations have shown that multilayered networks have potentially great power to learn complex discriminations. But this power was gained at an enormous cost; there was no natural rule for teaching such multilayered networks what they had to learn. For almost 20 years, no plausible learning rule had been proposed that would enable scientists to work with these multilayered networks; consequently, most scientists lost interest in using such networks as learning models.

Recently, however, three groups of researchers have independently discovered generalizations of the LMS rule that can plausibly be used with multilayered networks (Le Cun, 1985; Parker, 1985; Rumelhart et al., 1986). They have invented algorithms that propagate weight changes at the output layer back through successively earlier layers of unit connections. Multilayer nets that back-propagate learning changes have been demonstrated to learn many discriminations, such as parity, exclusive-or, and symmetry relationships. Since then, researchers have focused on exploring the information-processing potential of these LMS algorithms, trying to demonstrate their sufficiency for solving complex learning problems. Few researchers, however, have addressed the question of whether the LMS rule provides an empirically accurate account of how people learn. This is the ultimate aim of our research reported here. In developing an experimental program to begin to test the LMS rule as a component of human learning, we start by exploring its predictions for asymptotic behavior during probability learning. In doing this, we have begun to exploit a remarkable connection between animal and learning theory and adaptive network theory.

Connecting Models of Animal and Human Learning

As Sutton and Barto (1981) noted, the LMS rule is essentially identical to the Rescorla–Wagner (1972) model of associative learning in animals. This simple but powerful theory describing animals' learning in classical Pavlovian conditioning was presented by Rescorla and Wagner in the early 1970s (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). If in Equations 1 and 2 we let $V_i = w_{ij}$ set the training signal in the LMS rule, d_j, equal to λ when the US is present and to 0 otherwise and let $a_i = 1$ when CS_i is present and 0 otherwise, then Equation 5 of the LMS rule reduces to Equations 1 and 2 of the Rescorla–Wagner model. Curiously, adaptive network theorists have adopted the LMS rule because of its computational power, convergence properties, and generalizability to multilayered networks. Nonetheless, adaptive networks that implement the Rescorla–Wagner/LMS rule (henceforth LMS) can be viewed as a framework for modeling the emergent properties of complex configurations of the elementary associative processes found in animal conditioning.

Experiment 1

Because category learning is a central topic in cognitive psychology, we sought to evaluate adaptive networks as models of subjects' learning to classify stimuli into categories. This is also an elementary induction task for which the network model seems well suited. The focus of our evaluation is on comparing the performance of human learners with the expected asymptotic performance of a network that converges to the LMS solution (provided by Equation 5) to the learning problem.

In our experiment, university students, pretending to be medical diagnosticians, read the medical charts of 250 hypothetical patients, each described by the presence or absence of each of four symptoms (stomach cramps, discolored gums, etc.). The student diagnostician classified each patient as having one or the other of two fictitious diseases and received feedback regarding the correct diagnosis. During training, subjects learned which symptoms were more or less diagnostic of each disease. The symptoms, however, were not completely valid; that is, the cues were probabilistic, as in the studies of multiple-cue probability learning (Castellan, 1977) and the learning of fuzzy or ill-defined categories (Medin & Smith, 1984).

To develop a network model of subjects' performance in this task requires additional assumptions about (a) the network structure, (b) the representation of external events in the network, and (c) a mapping from network behavior to observable behaviors in people.

Figure 1 illustrates the simplest one-layer associative network that one could propose for this task. Each of the four symptoms is represented by an input node at the left. Because our task was presented to subjects as a forced choice between two mutually exclusive alternatives (i.e., every patient had

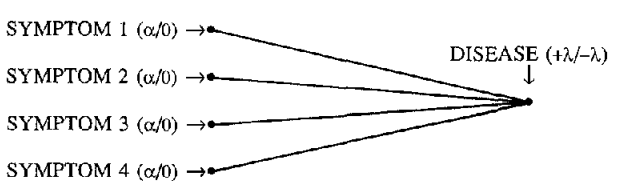

Figure 1. An adaptive network that learns to diagnose patterns of up to four symptoms as having one of two diseases.

one disease or the other), there is only one degree of freedom in the response. We can therefore represent the output (response) as a single node, assigning positive activation values on the output node to indicate increasing degrees of a preference for one disease, called Disease R, and the negative output values to indicate increasing degrees of preference for the alternative category, Disease C. The connection from symptom i to the output node has weight, w_i, reflecting the strength of evidence that the presence of symptom i provides toward a diagnosis of Disease R versus C. The four weights will be adjusted trial by trial according to the LMS rule in Equation 5.

Assume that presentation of a symptom pattern for a patient (a "trial") causes a pattern of activation (α if present or 0 if not present) over the four input units. The category (output) node is then activated in an amount equal to α times the sum of the weights from the presented symptoms to those category nodes (i.e., according to Equation 3). The activation reflects the model's expectation for Category R (versus C) given the symptom pattern. In the present situation the corrective feedback on a trial is the same regardless of the response. The amount of learning change produced by the feedback is larger the greater the difference between the current expectation and the training signal. The training signal provided to the output node (Figure 1) is the experimenter's feedback (after the subject's response) regarding the correct response. Assume that if Category (or Disease) R is the correct classification, then the feedback (d_j in Equation 5) will be set equal to $+\lambda$ on that trial; if the alternative category, C, is correct on a given trial, then the feedback will be $-\lambda$ for that trial.

Two of the parameters in the model are λ and α. These parameters can be shown to have only a multiplicative effect on the asymptotic weights and activation values. The weights and activations may therefore be considered as measured on a ratio scale; that is, they are unique to within a scalar multiple of λ/α. For convenience, let $\lambda = \alpha = 1$. Also assume a single learning rate parameter, β, for adjusting the weights. If a fixed set of training patterns is presented many times in random order to the learning model, the convergence properties of the LMS rule lead to expected asymptotic levels of the w_is, the symptom-to-disease associations, which are independent of β (see Appendix A and also Rescorla & Wagner, 1972; Parker, 1986). These weights, which are parameter-free predictions, are those that minimize the mean squared errors of classification achievable by the network for the training problem.

As in the Rescorla–Wagner (1972) formulation we assume that weights and activations will be mapped into decision (response) probabilities by a monotone transformation that preserves their ordinal relationships. Thus, higher output activations translate into higher probabilities of choosing Disease R over Disease C. Later we will propose a specific transformation to obtain choice probabilities. Other measures of associative strength are also possible. One we have used is to ask subjects to estimate directly the marginal probability that a patient with a particular symptom, for example, s_i, has one disease or the other, irrespective of his or her other symptoms. We will suppose that the greater the weight from s_i to the output node, w_i, the higher will be the subjects' estimates that the patient has Disease R rather than Disease C. Because of the symmetry in the reinforcement, positive weights will reflect a preference for Disease R, whereas negative weights will reflect a preference for Disease C. Because the weights in the model are specified on a ratio scale, the mappings of these weights onto subjects' estimates of disease likelihoods will be assumed to have only ordinal significance.

We will compare the predictions of the LMS rule in our category learning task with the predictions of three competing models of category learning (Estes, 1986): (a) *exemplar* models, which presume that the learner stores are 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); (b) *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 (c) *prototype* models, which presume that the learner abstracts the central tendency (model description) of each category and then classifies instances according to their similarity to this central prototype (e.g., Fried & Holyoak, 1984; Homa, Sterling, & Trepel, 1981).

When applied to our task in which subjects estimate the probability of each disease given each symptom, the models make one of two predictions. Exemplar models assume that subjects have access to all (or a random sample of) the exemplars presented during training. According to these models, subjects access all stored exemplars that contain a particular symptom and note the proportion of cases in which this symptom occurs with Disease R. Thus these models predict that subjects' estimates of the conditional probabilities will simply reflect the conditional symptom-to-category probabilities observed in the training sequence, a form of "probability matching." A pure prototype model that ignores variations in the overall base-rate frequencies of the diseases would predict that subjects' estimates of the probability of a disease given a symptom will simply reflect the relative likelihood of the symptom given the alternate diseases, namely,

$$\frac{P(s_i|C_1)}{P(s_i|C_1) + P(s_i|C_2)}.$$

In feature-frequency models, subjects are presumed to keep count of the number of times the symptom (feature) is associated with each category. If the model assumes that these counts are stored as the relative frequencies within a category, then the feature-frequency model makes predictions identical to prototype models, namely, that subjects' estimates should, incorrectly, reflect the relative likelihoods of the symptoms given the alternate categories. However, if the counts are stored directly, then the frequency model makes predictions identical to the exemplar models, namely, that subjects' estimates should reflect the objective conditional probabilities in the training sequence.

To generate differential predictions to compare for the LMS model and the alternative models, we need a learning task in which the ordinal relationships among the asymptotic weights differ from the ordinal relationships among either the objective posterior conditional probabilities of the categories given the features or the objective relative likelihoods of the features

given the categories. This is what we did in the following experiments. One way to arrange such a situation is to unbalance the overall frequencies of the two diseases so that one occurs far more often than the other. The question is whether people's probability estimates and choices will be more closely predicted by the LMS rule than by the exemplar, feature-frequency, or prototype models.

Method

Nineteen subjects were trained to classify medical charts of hypothetical patients into one of two mutually exclusive disease categories. Disease names were fictitious, but we refer to them as the rare (R) disease and the common (C) disease. Among the training exemplars, patients with the common disease were three times as frequent as patients with the rare disease. A patient chart consisted of one to four symptoms drawn from a set of four possible symptoms: bloody nose, stomach cramps, puffy eyes, and discolored gums.

Figure 2 (left side) shows the conditional probability of each of the four symptoms occurring in patients suffering from each of the two diseases. Each subject received a novel set of training patients that was generated during the experiment according to probabilistic procedure. First, each patient was randomly designated as suffering from either the rare disease ($p = .25$) or the common disease ($p = .75$). Second, given his or her disease, a patient's symptom chart was generated by choosing symptoms according to the following method (see Figure 2, left side): If the patient suffered from the rare disease, then with $p = .6$, the chart would include symptom s_1; with $p = .4$, symptom s_2; with $p = .3$, symptom s_3; and with $p = .2$, symptom s_4 (and analogously, but inversely, for patients suffering from the common disease). However, patients with no symptoms—henceforth, the "null" patients—were eliminated from the training sequence. This resulted in the true likelihoods of the symptoms given the diseases being slightly higher than indicated above (i.e., the probabilities of the symptoms given the rare disease were actually .69, .46, .35, and .23 for symptoms s_1 through s_4, respectively, and were analogous but inverse for the patients suffering from the common disease).

With the base rates of P (R) = .25 and P (C) = .75 and the probabilities in Figure 2 (left side), Bayes's theorem provides the conditional probability of the two diseases given the four symptoms considered separately (see Figure 2, right side). Note that for any single symptom the objective probability of the rare disease was always less than or equal to the probability of the common disease. Because a null patient was equally likely given either the rare or common disease, the computation of the posterior conditional probabilities is unaffected by the elimination of the null patients.

Each subject received 250 training trials of predicting diseases and receiving feedback on the cathode-ray tube (CRT) of a microcomputer. On each trial a new patient (corresponding to a list of symptoms) was presented on the CRT, and the subject had unlimited time to categorize the patient by pressing the R or C key on the microcomputer and then received feedback on the CRT about the correct diagnosis; the next patient's chart was then presented. The diseases were identified by fictitious names assigned to the rare or common disease in counterbalanced fashion across subjects. Subjects were told that there was no simple rule for making the diagnosis and that the order of presentation of the symptoms within a patient's chart was irrelevant. After training, subjects were tested by being asked to estimate directly the probability that a patient exhibiting a particular symptom was suffering from one or the other disease. For each symptom and disease, subjects were asked, "Of all the patients in the hospital exhibiting [symptom], what percent of those patients would you expect to suffer from [disease]?" They gave numerical estimates of $P(\mathrm{R}|s_i)$ and $P(\mathrm{C}|s_i)$ on a 0-to-1 scale for each of the four symptoms, for a total of eight estimates. Estimates were made by selecting 1 of 11 keys marked 0, .10, .20, . . . , .90, 1.00 in probability steps of .10. These estimates are the data of primary interest in this report.

Results and Predictions

To find asymptotic values for the association weights, one can derive an equation for the expected value of each weight at time t and then let t go to infinity (see Appendix A for details). As noted, the asymptotic weights depend only on the

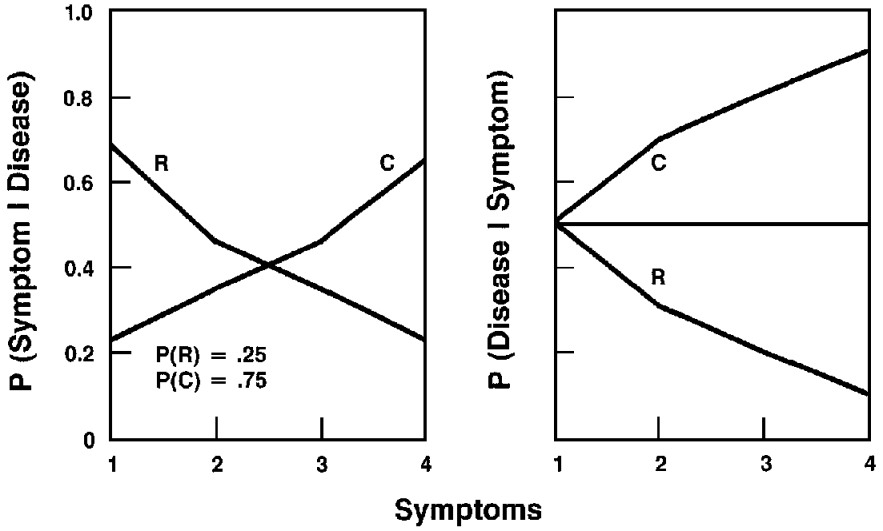

Figure 2. Experiment 1 design: On the left, the probabilities of each of the four symptoms occurring in patients suffering from each of the two diseases. The lower numbered symptoms were more typical of the rare (R) disease whereas the higher numbered symptoms were more typical of the common (C) disease. On the right, the conditional probabilities of each of the two diseases given the presence of each of the symptoms computed from the left-hand side by using the base rates and Bayes's theorem.

reinforcement probabilities in Figure 2 (right side) and not on the learning rate, β. The expected asymptotic association strengths are .430, −.043, −.306, and −.771 for w_1 through w_4, respectively, and these are plotted in Figure 3; these theoretical indices are to be compared with the objective conditional probabilities as well as the observed estimated conditional probabilities (also shown in Figure 3 for comparison).

The most striking difference between the objective probability measures in Figure 3 and the predicted associative weights in Figure 3 occurs for Symptom 1 (denoted s_1). This symptom was paired as often with the rare disease, R, as with the common disease, C; hence the objective conditional probability of R versus C is .5. However, the LMS rule predicts that s_1 will be associated more with the rare disease than with common disease, as indicated by a value of .43 for w_1.

This prediction of the LMS rule within the network model is understandable in light of the competitive nature of the learning algorithm. The asymptotic weights reflect the degree to which a symptom has been an informative and reliable predictor of a disease, relative to the predictive value of other symptoms that happen to be present at the same time. Although s_1 was paired equally often with the two diseases, it generally occurred in the company of other symptoms when the common disease was reinforced. On rare disease trials in which s_1 occurred, the other symptoms were far less likely to be present. Hence, w_1 was pushed more towards +1 (indicating the rare disease) than towards −1 (indicating the common disease). Whereas s_1 is not a better predictor of the common disease than the other symptoms, it is a relatively better predictor for the rare disease than are the others. It is the relative validity of a symptom for the two categories that determines its association with them in the LMS model.

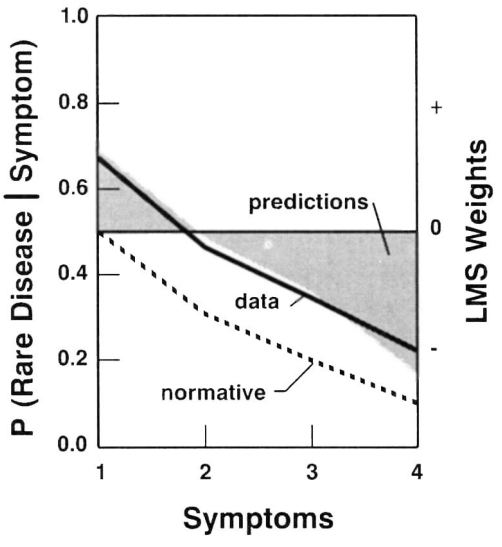

Figure 3. Results and predictions for Experiment 1. The objective probabilities of the rare disease given each of the symptoms are shown as a dashed line (with the scale along the left vertical axis); these also correspond to the predictions of exemplar and feature-frequency learning models. The predictions of the least mean square (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 0.5 on the left scale). The observed means of subjects' estimates of P(Rare Disease$|s_i$) are shown as a solid line (with the scale along the left vertical axis).

Observed estimates. Having described the model's predictions, we turn now to the data. Comparing the actual with the estimated conditional probabilities indicated that whereas subjects correctly learned the relative strengths of the conditional probabilities within a particular disease category, they appreciably overestimated the conditional probabilities of Disease R given each of the symptoms. Subjects' estimates of the probability of Disease R are graphed in Figure 3 for comparison with the predicted and objective values. Our preceding analyses suggested that the data for s_1 would be most critical for distinguishing between the models. As predicted by the LMS rule, 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; subjects' mean estimate of $P(\mathrm{R}|s_1) = .67$ was significantly greater than .50, $t(18) = 4.87$, $p < .0005$ (one-tailed), and over twice as great as their estimate of $P(\mathrm{C}|s_1)$. This simple result disconfirms the alternative models presented in the introduction, which predict that subjects' estimates will reflect the objective conditional probabilities observed in the training sequence.

Predicting classificatory responses. The measure used earlier (i.e., subjects' estimates of the conditional probabilities of the categories given the features) differs from that used in most studies of categorization performance. Our measure asks people to estimate probabilities on the basis of partial information, for instance, the presence of just a single symptom without further knowledge. On the other hand, the typical study asks subjects to decide on a category when given full information about an instance (e.g., the complete pattern of symptoms exhibited by a single patient). Therefore, it is instructive also to compare the ability of the models to predict subjects' classifications of patients when given complete symptom patterns. The 15 ($2^4 - 1$) possible patients or symptom patterns that could occur during the training trials are listed in Table 1. We refer to these as Patterns A through O.

The LMS model presumes that the total evidence or expectation for the rare disease is the sum of the association weights of the presented symptoms. To derive a choice probability from these activations or expectations requires a rule to map the activations of the output (category) node into a probability of choosing the rare disease in preference to the common disease. Such a mapping rule should have the property that negative numbers would be mapped to probabilities between 0 and .5, that 0 would be mapped to .5, and that positive numbers would be mapped to the range .5–1.

One of the simplest functions that satisfies this constraint is the logistic. If one lets S_k represent 1 of the 15 possible stimulus patterns, then one may write the logistic function as

$$p_k = P(\mathrm{R}|S_k) = \frac{1}{1 + e^{-\theta(O_k)}}, \qquad (7)$$

Table 1
Symptom Patterns Presented in Experiment 1

Pattern	Symptom s_1	s_2	s_3	s_4	P(R) Objective	LMS	Observed
A	0	0	0	1	.05	.08	.03
B	0	0	1	0	.18	.27	.15
C	0	0	1	1	.03	.03	.02
D	0	1	0	0	.34	.47	.25
E	0	1	0	1	.08	.07	.05
F	0	1	1	0	.25	.24	.27
G	0	1	1	1	.05	.03	.00
H	1	0	0	0	.67	.80	.66
I	1	0	0	1	.25	.25	.27
J	1	0	1	0	.56	.60	.55
K	1	0	1	1	.18	.11	.18
L	1	1	0	0	.76	.78	.73
M	1	1	0	1	.34	.24	.33
N	1	1	1	0	.67	.56	.53
O	1	1	1	1	.25	.10	.27

Note. P(R) = probability of rare disease; LMS = least mean squares.

where $P(\mathrm{R}|S_k)$ is the probability of responding "rare disease" given stimulus pattern S_k, O_k denotes the activation in the output node resulting from the presentation of S_k to the input nodes, and Θ is a (positive) scaling parameter. Equation 7 describes an ogive curve going from 0 when O_k is very negative, through .50 when $O_k = 0$, to 1 when O_k is very positive. Equation 7 is essentially Thurstone's (1927) law of comparative judgment and Hull's (1943) rule for converting reaction potential differences into choice probabilities. Recall that Equation 3 sets $O_k = \sum_{i=1}^{4} a_i w_i$, which is the sum of the weights of the presented symptoms.[2]

We fitted Equation 7 to the choice probabilities estimated for the 15 symptom patterns over the final block of 50 trials (out of 250). The Θ estimate was obtained by estimating the slope of log $[(1 - p_k)/p_k]$ plotted against the theoretical O_k values, excluding cases in which p_k was zero. The best estimate of Θ proved to be 3.2; the fit of Equation 7 with this Θ to subjects' choice probabilities for the 15 symptom patterns is shown in Table 1. The model's predictions correlate fairly well (.94) with the observed choice probabilities to specific symptom patterns with an average absolute discrepancy of .07. However, a chi-square goodness-of-fit statistic suggested that the fit could be much improved. Because many of the proportions in Table 1 are based on relatively small samples, they have large standard errors.

For comparison, we can estimate the expected performance of a probability learner who acquires all the relevant conditional probabilities of the two diseases for each symptom pattern and then chooses diseases for each symptom pattern to achieve probability matching (at the pattern-disease level). These probability predictions for this pattern-learning model are the objective values listed in Table 1. These probability-matching predictions correlate well (.99) with the data with an average absolute discrepancy of .03. The LMS model and the pattern-probability-matching models make fairly similar predictions (r = .95). In fact, the correlation between the two is so high that one could choose Θ in Equation 7 to fit the LMS model to pattern-probability matching (rather than the data). When one does this (yielding Θ = 2.8) and then proceeds to predict the data with that value of Θ, the correlation (.95) is by chance slightly better than before (due to p = 0 points left out of the earlier estimation procedures).

The implications of these results are twofold: (a) Both the LMS rule and the pattern-probability-learning model do fairly well in predicting subjects' choice behavior during the final stage of the training session, and (b) given that subjects' choice behavior is very close to that of a pattern-probability-matching learner, subjects appear to have learned the task as well as can be expected. This second point is important because, due to the probabilistic nature of the learning, subjects were trained only for a fixed number of trials rather than to some particular criterion performance.

Furthermore, as Table 1 shows, the LMS rule and pattern-probability model make very similar choice predictions across the symptom patterns. These choice data do not permit a strong preference between the pattern-probability-learning and the LMS model. The models do become distinguishable, however, when subjects are asked to estimate directly the likelihood of each disease given each symptom singly.

In comparing subjects' choices versus their likelihood ratings, it is particularly informative to examine the observed choice probabilities for Patterns H, D, B, and A, which are the patterns in which only one symptom was present (see Table 1). Subjects appeared to be aware of the conditional probabilities of the diseases given these patterns (as indicated by the probability matching of their choices). Note that the LMS model does not distinguish between the absence of information about symptoms and information about the absence of symptoms. Thus the model makes identical predictions for Pattern H (i.e., the presence of s_1 and the absence of symptoms s_2, s_3, and s_4) and for the case in which s_1 is present but no other information is given, that is, $P(\mathrm{R}|s_1)$. The actual conditional probability of the rare disease in these two cases is quite different, however. The actual probability that a patient with symptom s_1 and no other symptoms (e.g., Pattern H) has the rare disease is .67; in contrast, the probability that a patient who has symptom s_1 (but about whom no other information is known) also has the rare disease is .5, as shown

[2] Another perspective on Equation 7 is informative. Consider an alternative network representation of this problem with two output nodes, each connected to all four input nodes. One node—representing the rare disease—would be reinforced with 1 on rare disease trials and with 0 on alternative trials. The other node—representing the common disease—would be similarly reinforced except that it would receive the reinforcement of 1 on common-disease trials. Such a network would be conceptually equivalent to simultaneously "conditioning" the symptoms to the alternative diseases to various extents, according to the Rescorla–Wagner (1972) model. If the activation levels on the two output nodes are transformed by an exponential function, that is, $O_{\mathrm{R}} = e^{\Theta \sum_{i=1}^{4} w_{i\mathrm{R}} a_i}$, where O_{R} is the activation in the rare output node and $w_{i\mathrm{R}}$ is the weight from Symptom i to the rare node, and then one applies ratio response rule to get choice probabilities, with $P(\mathrm{R}|S_k) = \frac{O_{i\mathrm{R}}}{(O_{i\mathrm{R}} + O_{i\mathrm{C}})}$, one would arrive at exactly the same expression as Equation 7.

in Figure 2 (right side). The difference between these two calculations (.67 vs. .50) comes from the impact of the information about the absence of the other symptoms in Pattern H. Because the observed $P(\mathrm{R}|\text{Pattern H}) = .66$ is close to the subjects' judgments about the $P(\mathrm{R}|s_1) = .67$, it is tempting to suggest that perhaps subjects were basing their estimates of $P(\mathrm{R}|s_i)$ on their estimates of the conditional probability of the pattern that contained only s_i and no other symptom (e.g., Pattern H). An examination of the other single-symptom patterns, however, revealed that the observed choice probabilities were .25, .15, and .03 for the patterns containing only symptoms s_2, s_3, and s_4, respectively, whereas the observed direct ratings of $P(\mathrm{R}|s_i)$ for these symptoms were .46, .34, and .22, respectively. Thus, although this confusion may contribute to subjects' judgments, it does not provide a fully satisfactory account of the data. Moreover, evidence collected in later experiments suggests that nearly all subjects understood the difference between these two cases, s_i alone versus s_i with the s_j absent.

Discussion

The primary results of this experiment, that is, the probability estimate differences shown in Figure 3, confirm the predictions of the LMS rule within this network model. It is important that 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 regardless of what symptoms the patient presented.

This result suggests that the subjects fell prey to a 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 many results in research on judgment: 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). When answering questions such as "What is the probability that Object A belongs to Class B?", people often resort to a representativeness heuristic in which their judgment reflects the degree to which Object A resembles a prototype of Class B objects (Tversky & Kahneman, 1982). For example, in estimating the probability that a particular student is a computer science major in a classroom known to be 80% English majors, people base their predictions largely on the degree to which the personality characteristics of the student are representative of their stereotypes of computer wizards, thus neglecting the influence that an 80% base rate should have (Kahneman & Tversky, 1973). Most studies demonstrating such neglect of base rate in classification judgments have used natural categories with familiar prototypes (e.g., feminists and engineers), and base-rate information has generally been presented to subjects as abstract numerical information (Tversky & Kahneman, 1982). Base-rate neglect was demonstrated in an experiment in which 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.

One might try to explain our results in Figure 3 by supposing that subjects completely ignored base rates of the two categories in making their judgments. For instance, we earlier indicated how prototype and feature-frequency models could be interpreted to be insensitive to base rates of the two categories. But this explanation fails because if subjects had been ignoring base rates, then they should have judged symptoms s_1 and s_2 to be as diagnostic of the rare disease as they judged symptoms s_3 and s_4 to be of the common disease. (See the symmetry of Figure 2, left side, around .5.) But as Figure 3 shows, this pattern was not obtained. Only s_1 was judged to be a significantly stronger predictor of the rare disease than the common disease.

Though subjects' probability estimates reflected less influence because of base rates than objectively required, the estimates nonetheless show definite sensitivity to the differing base rates. These results are consistent with current judgment studies that suggest that in most situations base-rate information is not ignored, only underused (Borgida & Brekke, 1981; Kassin, 1979). Alternative category-learning models, which predict either total neglect of base rates or full normative use of base-rate information, do not provide an adequate account of the data. The LMS rule, however, correctly predicts that in this situation only s_1 will be perceived as stronger evidence for the rare disease; the other symptoms are predicted to be stronger evidence for the common disease.

This LMS-network model views the base-rate neglect observed here as the outcome of a learning process (of adjusting weights according to Equation 5). Other demonstrations of base-rate neglect, however, have involved only flawed judgment processes, not learning. Thus Bar-Hillel (1980) showed that even when given direct numerical values for base rates and the probability of the features given the categories, subjects still underused the base-rate information, as though their calculations were flawed. Conceivably, our subjects may have learned veridical probabilities of the diseases and the disease-to-symptom associations but then miscalculated (as did Bar-Hillel's subjects) in directly estimating the probabilities of the rare disease given each symptom.

Such a possibility could be checked by asking subjects to estimate the base rate, $P(\mathrm{R})$, and the conditional probabilities of the symptoms given each disease as well as the disease-given-symptom conditionals (which we obtained). In its simplest form the network model would interpret the strength of association from a disease to a symptom to be identical to the association from that symptom to that disease. Unpublished data collected by Gluck (1984) suggested this is true: In a category-learning experiment that unconfounded $P(\text{Category C}|\text{Feature X})$ from $P(\text{Feature X}|\text{Category C})$, subjects who had been trained to discriminate different levels of $P(\text{Category C}|\text{Feature X})$ completely transferred these values over when they later estimated the reverse probabilities, $P(\text{Feature X}|\text{Category C})$. They behaved as though the latter quantity were equivalent to the former quantity, on which they had been trained.

Experiment 2

Many learning phenomena—including overshadowing, blocking, and the effect of intertrial presentations of the unconditioned stimulus—indicate that animals learn about the contingency or informational value of the CS (Prokasy, 1965; Rescorla, 1968) rather than learn simply about the contiguity of the CS and US (Hull, 1943; Spence, 1956). The importance of the Rescorla–Wagner (1972) model is that it posits a single associative process that accounts for the role of these informational variables. In the second experiment we sought to test directly one of the basic predictions of this LMS-network model, namely, that different cues compete to be the more valid predictor of an outcome.

If subjects in the category-learning experiment are basing their probability judgments on subjective association strengths learned according to the LMS model, then one might expect to find analogs of the competitive learning effects that occur in classical conditioning experiments. For example, to the extent that a stimulus cue is redundant with a stronger or more valid cue in predicting a US, the model expects that the associative strength of that cue will be greatly attenuated (Wagner, 1969). Analogously, in the category-learning paradigm one should be able to attenuate the apparent diagnosticity of symptom s_1 by making one of the other symptoms a truly reliable and strong predictor of the rare disease. Adding a strong predictor of the rare disease should greatly attenuate a subject's erroneous belief that s_1 is itself strong evidence for the rare disease. The extent of this attenuation should increase the more often the valid predictor cooccurs with s_1. Thus in our category-learning paradigm we expected to attenuate the apparent diagnosticity of symptom s_1 (found in Experiment 1) by making one of the other symptoms a truly reliable and strong predictor of the rare disease. To test this prediction we designed Experiment 2 to be identical to Experiment 1 except that symptoms s_2 and s_3 were more valid predictors of the rare and common diseases, respectively. In particular, the likelihood of s_2 given the rare disease was increased, and its likelihood given the common disease was decreased: s_2 was present with $p = .9$ in rare disease patients and with $p = .1$ in common disease patients. Similarly, but conversely, the likelihood of s_3 given the rare disease was decreased, and its likelihood given the common disease was increased: s_3 occurred with $p = .1$ in rare disease patients and with $p = .9$ in common disease patients.

A comparison of the critical conditional probabilities used in the two experiments is shown in Figure 4. This illustrates the actual conditional probability differences for the two symptoms, s_1 and s_4, whose objective relationships remained unchanged from Experiment 1 to Experiment 2 (see Figure 3). Symptom s_1 is still equally diagnostic of the two diseases (and hence the objective value of $P(R|s_1)$ is .5). A model that assumes that subjects are independently learning the pairwise contingencies of the symptoms and diseases predicts that the manipulation of predictiveness of s_2 and s_3 should not influence the judgments for symptoms s_1 and s_4.

The LMS rule, however, is a competitive learning algorithm in which co-occurring cues compete to predict the outcome, and those cues that are the most valid and nonredundant are favored for strengthening. Applying the LMS rule to the design in Experiment 2 (see Appendix A), we derive asymptotic association strengths of .18, .76, −.94, and −.20, for s_1 through s_4, respectively.

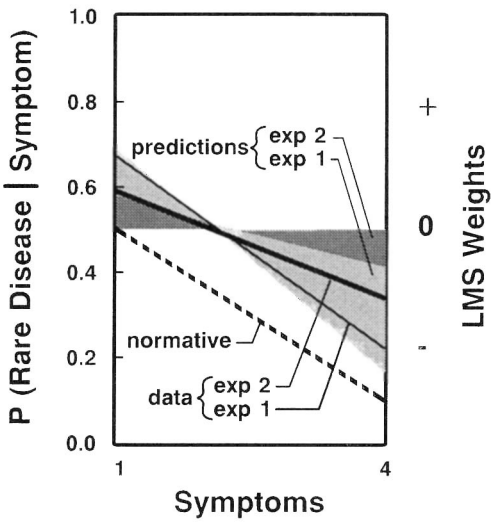

Figure 4. Results and predictions for Experiment (exp) 2. The objective probabilities of the rare disease given Symptoms s_1 and s_4 are shown as a dashed line (with the scale along the left vertical axis); these also correspond to the predictions of exemplar and feature-frequency learning models. The predictions of the least mean squares (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 0.5 on the left scale). The observed means of subjects' estimates of P(rare Disease$|s_i$) are shown as a solid line (with the scale along the left vertical axis).

The weights are graphed in Figure 4, along with the analogous measures from the first experiment (taken from Figure 3). As anticipated, the LMS rule expects considerable attenuation of the association strength between s_1 and Disease R, and between s_4 and Disease C.

Method

Thirty-four new volunteers from Stanford University's introductory psychology course served as subjects. The procedure in this experiment was identical to that of Experiment 1. The only difference was in the probability of symptoms s_2 and s_3 being present in patients having Disease C versus Disease R (see Figure 4). Following 250 training trials, the subjects gave numerical estimates of the probability of each disease given each symptom considered singly.

Results

As predicted by the LMS rule, the results in Experiment 2 showed that the apparent diagnosticity of symptom s_1 for the rare disease was significantly attenuated by introducing greater validity for symptom s_2 toward the rare disease. The

direct estimate of the conditional probabilities of the rare disease given symptom s_1 (without information of other symptoms) decreased from .67 (in Experiment 1) to .54 in Experiment 2. This was a significant decrease, as the model predicted, $t(51) = -1.93$, $p < .05$, (one-tailed). The model also predicted that symptom s_4 would have its association with the common disease attenuated. The direct estimate of the conditional probabilities of the rare disease given symptom s_4 increased from .24 (in Experiment 1) to .34 in Experiment 2. This was also a significant increase, as the model predicted, $t(51) = -2.04$, $p < .025$ (one-tailed).

Predicting Classificatory Responses

Again one may compare the predicted disease-choice proportions with specific symptom patterns obtained over the last 50 training trials. Recall that the observed proportions have a large standard error because of the small sample sizes. Using Equation 8 as in Experiment 1, we estimated $\Theta = 4.6$. Using the symbols in Table 1 to represent the symptom patterns, we plotted the observed proportions of rare disease choices to the patterns in Figure 5 against the proportions predicted by the LMS rule.

From the scatter of the data points along the diagonal, one sees that the correlations with the observed proportions are quite high: $r = .97$ for the LMS model (average discrepancy was .09) and also $r = .99$ for the probability-matching predictions (average discrepancy was .07). The goodness of fit, however, was significant for both models: $\chi^2(14) = 53$, $p < .001$, for the LMS predictions and $\chi^2(14) = 91$, $p < .001$, for the probability-matching predictions. Once again the predictions of the two models are highly correlated ($r = .97$) with each other.

Discussion

The results of Experiment 2 confirm a critical prediction of the LMS rule for learning. The apparent diagnosticity of symptom s_1 in Experiment 1 was significantly attenuated in Experiment 2 when s_2 became more diagnostic. It is as though symptom s_1 "lost its punch," or its claim 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 edge over other learning rules. The rule implies that people do not learn sets of independent cue–outcome associations, wherein the contingent correlation for each cue to each outcome can be calculated separately. According to the LMS rule, independent cue–outcome correlations have little behavioral impact. The key to associative strengthening of a cue is its validity or diagnosticity relative to other cues present in the situation.

It was this key variable, the relative validity, that was changed from Experiment 1 to Experiment 2. Among the cues for the rare disease in Experiment 1, symptom s_1 was relatively more valid. But in Experiment 2, symptoms s_2 and s_3 were made more valid, so that symptoms s_1 and s_4 lost their former associative differential, despite maintaining the same correlation with the diseases as they had in Experiment 1. The downgrading of these symptoms in Experiment 2 is analogous to the overshadowing of a stimulus by a more salient or more diagnostic stimulus in compound-cue classical conditioning. Indeed, our results echo conditioning phenomena such as blocking, overshadowing, and low CS–US correlations, which were originally offered as evidence for the competitive nature of cue–outcome associative learning.

Experiment 3

The closest competitor to the LMS rule is the learning rule that implies asymptotic probability matching at the level of whole patterns. Estes (1959) long ago formulated two models that achieved different versions of probability matching. He called these the *component* model and the *pattern* model. If one coordinates his terms to our experiment, a component (or stimulus element) would correspond to a single symptom, whereas a pattern would correspond to the entire configuration of symptoms presented as a patient. Thus our experiment

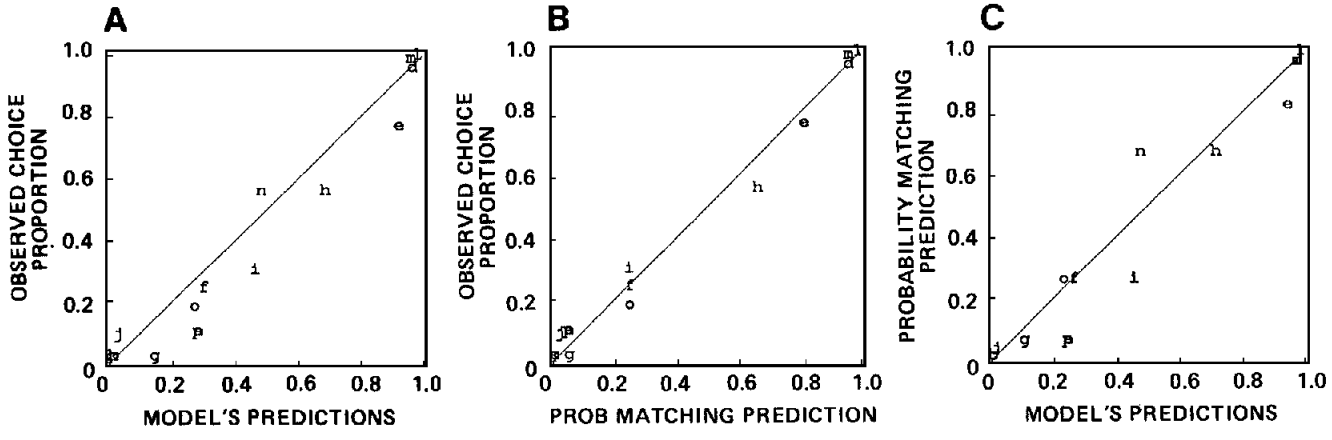

Figure 5. Scatter plots of (A) least mean squares (LMS) rule predictions versus observed choice probabilities for each of the 15 symptom patterns, (B) context model predictions versus observed choice probabilities, and (C) LMS rule predictions versus context model predictions for Experiment 2.

has four stimulus components but 15 stimulus patterns ($2^4 - 1$). The component model is identical to what Estes (1985) later called the feature-frequency model.

Given these stimulus identifications, one can then write linear difference equations corresponding to the trial-by-trial impact of Disease C versus R feedback trials in changing the associative probability of that stimulus (component or pattern) to the disease categories. If the changes in associative probabilities on the two kinds of disease trials are symmetric, then the long-term impact of a training schedule is to produce stimulus-to-category associative probabilities that match the objective probabilities of each disease category given that stimulus. When one coordinates the theoretical stimuli with the component symptoms, then the model expects the symptom-to-disease probabilities to be matched. According to this component identification, the proportion of Rare disease responses to a compound of symptoms (e.g., s_1 and s_3 and s_4) would be calculated by averaging the associative probability of each presented symptom toward the rare disease. But this component-matching model has been shown to fail to predict any aspect of the results.

Alternatively, one may identify the stimuli of the theory with complete patterns of symptoms and suppose that these become associated with the disease categories as intact units. Estes (1959) called this the pattern model. By this identification, then, each of the 15 symptom patterns in the experiment would tend asymptotically to become associated with Disease R with a probability equal to the likelihood that when that pattern occurred the patient had Disease R, that is, would match the actual P(R|pattern). Given the experimental base rates of the categories and the independent conditional probabilities of the symptoms given each disease, one can calculate these objective probabilities. It was those objective pattern probabilities that compared so favorably with the observed choice proportions in Table 1.

A curious contrast among the LMS model, Estes's (1959) two models, and the data may be noted here. Estes's (1959) component model predicts the data rather poorly, whereas the pattern model fares far better, especially with the choice proportions. On the other hand, the LMS rule that identifies the stimuli with the four symptoms (the components) does quite well in predicting the asymptotic choice proportions to the patterns. In fact, in this respect the LMS model with stimuli identified as components nearly mimics the Estes (1959) pattern model in its asymptotic choice predictions. The main discrimination between the two models occurred in the subjects' direct estimates of the probabilities of the two diseases given each of the symptoms singly.

In terms of Estes's (1959) pattern model, one can imagine that subjects derive such single symptom estimates in one of two ways. The first way would be to report the (weighted) average probability of Disease R as pooled over the eight (or 2^3) patterns that contained symptom s_1. This is equivalent to the Bayesian probability of Disease R conditional on appearance of symptom s_1 and is the measure plotted in Figures 2 (left side) and 3. Such predictions fail in systematic ways (see Figure 3). The second way is to set the probability of Disease R (given only knowledge of cue s_i) equal to the theory's probability of Disease R (given the single pattern with s_i present and the other symptoms absent). Clearly, however, these estimates are also less accurate than those of the LMS-network model. Thus neither rule for calculating direct estimates from the pattern model has proved successful.

In the next experiment, we sought an arrangement that might test more directly the LMS model against Estes's (1959) pattern model. The design chosen hinges on a difference in the way the component and pattern models treat absence of a symptom (patient does not have a runny nose) in contrast to presence of an opposite symptom (patient has a stuffy nose). Imagine that four pairs of opposing or mutually exclusive symptoms (e.g., runny nose and stuffy nose) are defined and that they are four binary dimensions.[3] Each patient is then defined as having one or the other value on each of the four dimensions. One can then carry out the basic learning design in Experiment 1. However, when the stimulus schedule calls for absence of symptom s_i, one replaces that by the mutually exclusive, opponent symptom, called s_i^*.

Interestingly, Estes's (1959) pattern model treats these two situations as practically identical, because the model treats patterns as unanalyzable wholes. As long as the 16 different patterns in Experiment 3 have the same probabilistic associations with the diseases as did the corresponding patterns in Experiment 1, the pattern model predicts that the same behavioral profile should be learned.

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 Figure 6. Network A represents each opponent symptom pair as a single input node that receives activation of $+\alpha$ or $-\alpha$, depending on which member of the opponent pair is presented on a given trial (patient); we refer to this as the *four-component* model. 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 treat the experiment in terms of Network A, because it would compute a perfect negative correlation between each s_i and s_i^*; hence, an eight-variable regression model would collapse to a four-variable model.

Both of these network representations are plausible and, indeed, correspond to different stances in the connectionist literature. One obvious implication of Network A is strong symmetry in activation for a given pattern and its complement (obtained by using the alternate values in the pattern). For example, if Pattern I (from Table 1) is $s_1 - s_2^* - s_3^* - s_4$ and yields output activation G, then its complementary pattern (Pattern F from Table 1), $s_1^* - s_2 - s_3 - s_4^*$, should yield output $-G$. There are eight such complementary pairs of patterns, so strong tests of this symmetry prediction from Network A are available. Incidentally, if subjects match probabilities in

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

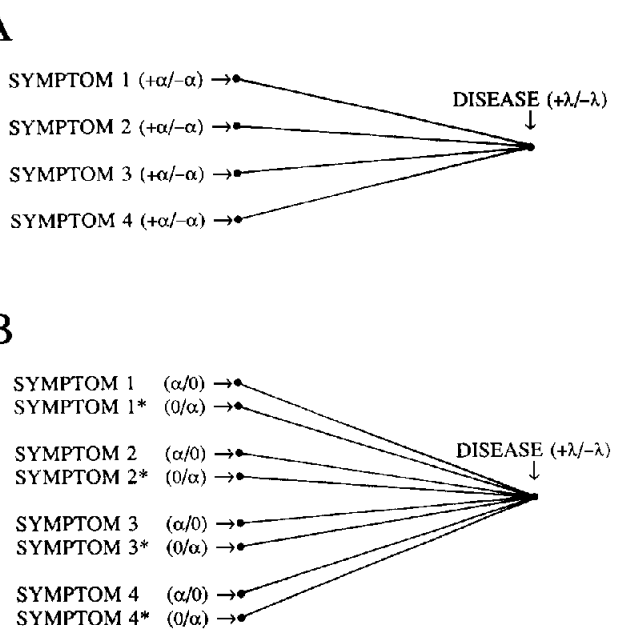

Figure 6. (A) A four-component network for classifying the stimuli from Experiment 3 that represents each opponent symptom pair as a single input node that receives activation of $+\alpha$ or $-\alpha$, depending on which member of the opponent pair is presented on a given trial (patient). (B) An eight-component network for classifying the stimuli from Experiment 3 that 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.

their diagnoses of symptom patterns (as they did in Experiments 1 and 2), then this will violate the four-component model's predictions for complementary patterns.

By a related line of reasoning, similar predictions follow from Network A about subjects' estimates of $P(\mathrm{R}|s_i)$. The four-component model predicts that subjects' estimates of $P(\mathrm{R}|s_i)$ and $P(\mathrm{R}|\text{not } s_i)$ will sum to one. These predictions provide further tests of the four-component model. On the other hand, Network B, treating the eight symptoms as distinct, makes no such strong predictions.

Turning first to the predictions of Model B, Figure 7 shows the predicted weights for the eight-component model. The derivation of these asymptotic weights is slightly different from that in Experiments 1 and 2; see Appendix B for details. Also shown are the probability-matching predictions of Estes's (1959) pattern model and, for comparison, the association weights predicted by the LMS model for Experiment 1.

Note several features of these predictions. First, the pattern model's predictions for the positive symptoms (s_1, s_2, s_3, and s_4) in Experiment 3 (see Figure 7) are identical to what they were in Experiment 1 (see Figure 3). Second, the predictions of the LMS model of the estimates of $P(\mathrm{R}|s_i)$ for symptoms s_1 through s_4 differ mainly by being less extreme (i.e., closer to .5) in Experiment 3 than they were in Experiment 1. Thus, compared with Experiment 1, symptom s_1 in Experiment 3 should appear to be less diagnostic of the rare disease, whereas symptoms s_3 and s_4 should appear less diagnostic of the common disease. In fact, the LMS model expects a general regression toward .50:.50 in the estimates of disease probabilities in Experiment 3. One reason for this regression is that in Experiment 3, a symptom must always appear in the company of three other symptoms; thus it must always share strengthening increments.

Third, these two models differ in their predictions for symptom s_4^*. The pattern model implies that s_4^* will be associated with common disease, whereas the LMS model predicts that s_4^* will be more associated with the rare disease.

The differing predictions from the two network models along with the predictions of Estes's (1959) pattern model were so compelling that we conducted Experiment 3 to see which ones would be closer to the results.

Method

As noted earlier, the statistical design and procedures of Experiment 1 (see Figure 2) were repeated,[4] except that each present or absent symptom (e.g., fever or no fever) was replaced by two mutually exclusive features (e.g., diarrhea or constipation), one of which was always present for each patient. Each patient had four symptoms. Thirty-six college-student subjects 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), providing a total of 16 estimates.

Results

Direct likelihood ratings. The primary results are subjects' estimates of the likelihood of Disease R versus Disease C. Figure 7 shows the observed 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 test the four-component model. As noted earlier, Network Structure A, which uses $+1/-1$ on input nodes to represent the pairs of opponent symptoms, predicts that judgments of $P(\mathrm{R}|s_i)$ should be symmetric around .5 with $P(\mathrm{R}|s_i^*)$. The data, however, are not consistent with this prediction; this is most clearly evident in subjects' estimates of $P(\mathrm{R}|s_3)$ and $P(\mathrm{R}|s_3^*)$, which were both significantly below .5, violating the symmetry prediction of the four-component LMS model.

For the eight-component model (Network B), the LMS rule predicts that 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. The estimated $P(\mathrm{R}|s_1)$ was significantly greater than .5, $t(35) = 5.46$, $p < .0005$ (one-tailed).

Second, Network B predicted that the opponent symptom, s_4^*, which is complementary to s_4, would be rated as diagnostic of the rare disease. Indeed, the observed estimates of $P(\mathrm{R}|s_4^*)$ are moderately greater than .5, $t(35) = 1.48$, $p < .10$ (one-tailed). The apparent diagnosticity of symptom s_4^* for the rare

[4] Actually, they differed slightly in that Experiment 1 excluded the pattern in which a patient had none of the four symptoms, whereas Experiment 3 had no such case.

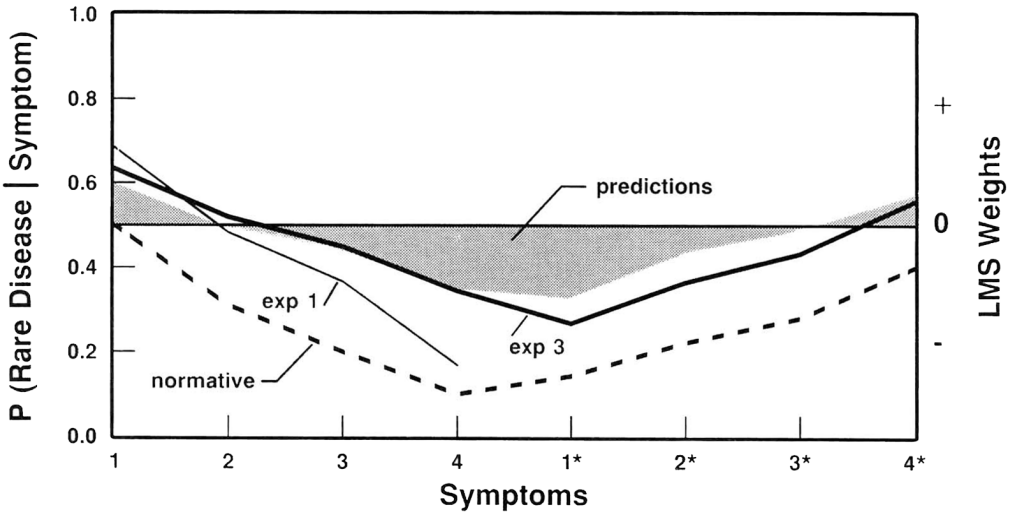

Figure 7. Results and predictions for Experiment (exp) 3. The objective probability of the rare disease given each of the symptoms is shown as a dashed line (with the scale along the left vertical axis); these also correspond to the predictions of exemplar and feature-frequency learning models. The predictions of the least mean squares (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 $P(\text{Rare Disease}|s_i)$ are shown as a solid line (with the scale along the left vertical axis). The data from Experiment 1 are also shown for comparison. Note that the objective conditional probabilities for Symptoms 1 through 4 were unchanged from Experiment 1 to Experiment 3.

disease violates the probability-matching prediction (see Figure 7) of the pattern model. Also, symptom s_4^* was significantly less diagnostic of the rare disease than was symptom s_1, $t(35) = 3.01$, $p < .005$ (one-tailed); the Network B model predicted a slight trend in this direction.

Third, comparing the estimates in Experiment 3 with those in Experiment 1, we expected the probability estimates for symptoms s_1, s_3, and s_4 to shrink toward .5, reflecting less competitive dominance. Pooling symptoms s_3 and s_4, there is a significant shrinkage of the observed estimate (toward .5) in Experiment 3 versus 1, $t(53) = 2.93$, $p < .005$ (one-tailed). The change for s_1 in Experiment 3 versus Experiment 1 is in the predicted direction but not statistically significant ($t = .95$, $p > .10$).

Fourth, the slight change in symptom s_2 (from below .5 to above .5) was not in the expected direction; however, the change was very small and did not differ reliably from .5 or from the small difference observed in Experiment 1.

Choice proportions. As before, one may examine the models' predictions of the asymptotic proportion of rare-disease choices for each of the 16 symptoms patterns, where Pattern P represents all-Os pattern that was missing in Table 1. Figure 8A compares the observed proportions with those predicted by Network Model B and Equation 7 with $\Theta =$

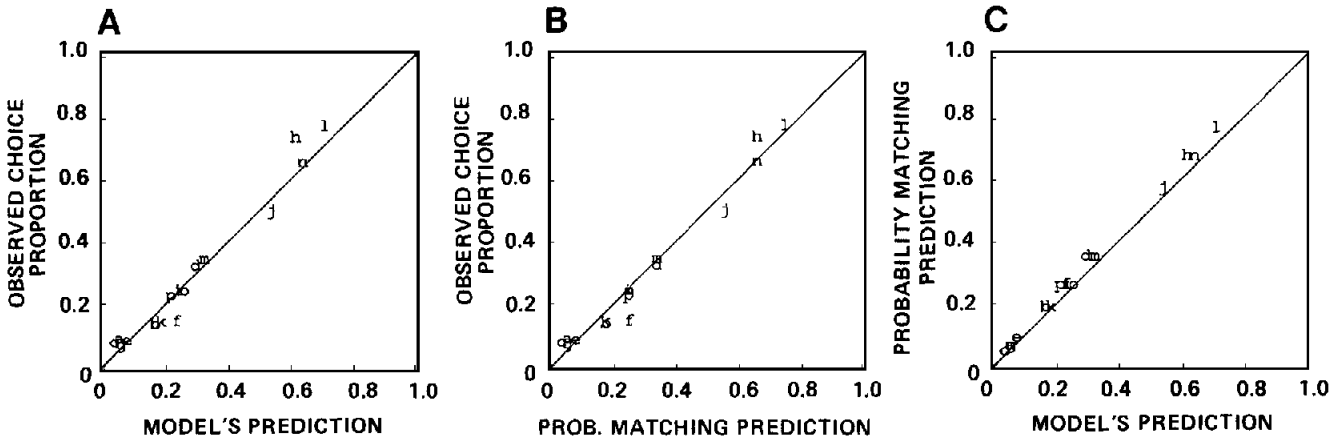

Figure 8. Scatter plots of (A) eight-component model predictions versus observed choice probabilities for each of the 16 symptom patterns, (B) pattern model predictions versus observed choice probabilities, and (C) eight-component model predictions versus pattern predictions for Experiment 3.

3.12, Figure 8B compares them with those predicted by probability matching, and Figure 8C compares Network Model B's predictions with those of probability matching. As in the earlier experiments, both the probability-matching and Network Model B's predictions correlate very highly with the observed proportions, especially given their large standard errors. The correlations with the observed proportions are .98 for the LMS model (average discrepancy was .08) and also .98 for the probability-matching predictions (average discrepancy was .03). The goodness-of-fit statistics are $\chi^2(15) = 22.4$, $p > .08$, for the LMS predictions and $\chi^2(15) = 22.6$, $p > .08$, for the probability-matching predictions. (Once again the predictions of the two models are highly correlated [.99] with one another.)

Turning back to the four-component model, one recalls that this model predicts that estimates of $P(R|S)$ and $P(R|S^*)$ will sum to 1, in contrast to the objective values of these probabilities. Contrary to this model's predictions, subjects' choice proportions were very close to the objective probabilities. Furthermore, their choice probabilities clearly violated the aforementioned equality; for example, in the two opponent patterns considered earlier, subjects' estimates of $P(R|\text{Pattern O})$ and $P(R|\text{Pattern P})$ summed to .45 rather than 1, a clear violation of the predictions of the four-component model.

Discussion

The probability ratings appear to support the eight-component LMS model (Network B), as opposed to the four-component LMS model (Network A) and the pattern model. Both the eight-component LMS model and the pattern model predict a V-shaped graph over the eight symptoms, but the LMS model predicts a V that rises above the .5 baseline at the two ends. Also, the pattern model expects the probability estimates in Experiment 3 for symptoms 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 two symptom lists deviate significantly from one another.

An advocate for the pattern model might argue that because Experiment 3 involved more cues (eight vs. four), the contingencies would be learned more slowly, so that subjects would not have reached asymptote in the 250 trials of training. Thus their likelihood judgments might be expected to be less extreme than predicted by probability matching. While this "regression to chance" argument has some merit, we note that it does not account for the salient results predicted by the LMS model, namely, the significant overexpectation of the rare disease given symptoms s_1 and s_4^*.

General Discussion

We used adaptive network theory to evaluate learning rules to describe human probabilistic category learning. As noted, to specify an adaptive network model, the theorist must specify a network architecture, identify a coding of environmental stimuli and responses in that network, and specify a rule for changing weights adaptively so that the system can learn to assign each input pattern to its appropriate output pattern. The network architecture we have adopted is the simplest possible for a forced choice task, namely, one layer of N distinct input units (one per physical stimulus) feeding activation directly into one output unit corresponding to the two output (disease or response) categories. We have tested this simple model and eschewed more complex theoretical options, such as multiple layers of association units "hidden" between the input and output layers, recurrent feedback loops, or multiple-unit representations of each symptom (for the range of possibilities, see Rumelhart & McClelland, 1986). Surprisingly, with just the simplest network, the optimal LMS weights delivered successful predictions in the three experiments reported here.

To briefly recap those predictions, the learning conditions arranged in Experiment 1 were expected to cause subjects to believe that symptom s_1 was significantly more diagnostic of the rare than the common disease. This prediction was confirmed in contradistinction to alternative theories that predicted that subjects' expectations would match the probabilities of the two diseases given each symptom. By converting theoretical strengths into choice probabilities by using the logistic function in Equation 8, the LMS model's predictions were also reasonably close to the pattern of rare-disease choice proportions as asymptote given each of the 15 patterns of symptoms.

The illusory diagnosticity of symptom s_1 for the rare disease in Experiment 1 depended on the fact that symptom s_1 was indeed relatively more diagnostic of the rare disease than were any other symptoms that might have been present; the LMS learning algorithm exploits this relative validity of a cue compared with that of its copresent competitors. This analysis suggests that the illusion of s_1's diagnosticity could be greatly reduced by increasing the validity of one of the other symptoms for predicting the rare disease. Conditions similar to this were arranged in Experiment 2, and the behavioral outcomes were as predicted.

In Experiment 3, we tested the LMS model against Estes's (1959) pattern model, which had closely predicted choice proportions to symptom patterns in Experiments 1 and 2. The pattern model implies that the same behavior profile would be learned when patient-symptom patterns were defined by the presence or absence of four symptoms as compared with the presence of a symptom or its opponent symptom. In contrast, Network B expected the opponent-symptom experiment to lead to quite different asymptotic outcomes. Given contingencies comparable to those in Experiment 1, the use of opponent symptoms in Experiment 3 was expected to diminish the extremity of strength differentials to the two diseases for symptoms s_1, s_2, s_3, and s_4, while leading to a reliable overexpectation of the rare disease for the complementary symptom, s_4^*. Such results were obtained in estimated disease probabilities, and they differed from those predicted by the pattern model.

One of the pleasant surprises was that the LMS association strengths can be easily converted into choice probabilities (by Equation 7), which provided a close fit to observed disease choices for symptom patterns. A curiosity is that the logistic transformation of the LMS weights can be chosen to be highly

correlated with the objective probabilities of the rare disease given each stimulus pattern. In fact, the mimicry is so accurate that one would be tempted to estimate Θ by fitting the LMS weights to the objective probabilities of the rare disease given each pattern; if one did so and then predicted the observed choice proportions, those predictions would then be virtually parameter-free, in that no data would be used to estimate Θ.

Note again the peculiar fact that the LMS model that identifies the stimulus units with individual symptoms delivers predictions of near-matching choice probabilities to patterns of symptoms, which can be achieved within Estes's (1959) theory only by identifying the stimulus units with complete symptom patterns. As noted before, the two models differ in their predictions of subjects' direct likelihood estimates of disease probabilities given knowledge of only one symptom at a time (e.g., see Figures 3, 4, and 7).

Related Work

Our experiments provide evidence for the competitive nature of learning, whereby cues compete for association strength according to their relative validity for predicting categories. Two recent results provide similar support for this competitive rule in studies of human cue-correlation learning.

One supportive result came in an experiment by Medin and Edelson (1988), who demonstrated an extreme form of learned neglect of base rate. They presented Symptom Pair AB with Disease 1 three times as often as Symptom Pair AC with Disease 2. As expected, this led to a majority of Disease 1 choices when subjects were tested with the ambiguous A cue. More important, a conflict test on the novel pattern BC yielded a surprising majority of Disease 2 choices, thus reversing the direction of the 3-to-1 base rate. As Medin and Edelson noted, the LMS learning rule accounts for this stronger association of Cue C to Disease 2 than of Cue B to Disease 1. Because the 3:1 base rate causes A to become predominantly associated with Disease 1, it increases the predictability of Disease 2 to AC. In such circumstances, the LMS rule implies that Cue B will be relatively more blocked than Cue C in acquiring their respective associations, so that Cue C will dominate B in the BC conflict test. And this was the paradoxical reversal of base rate that was to be explained.

In related research by MacMillan (1987), subjects were found to keep track of the relative frequency of a feature (cue) within a category only if it was somewhat diagnostic for the presence versus absence of that category. For example, if 70% of all patients with a target disease had stomach cramps and so did 70% of patients without the disease, subjects came to ignore this symptom because it was irrelevant for diagnosing the disease. More important, when asked to estimate the proportion of disease patients who had this symptom, subjects grossly underestimated the actual relative frequency. MacMillan showed that this neglect of relative frequency of invalid features develops over training as subjects learn the irrelevance of these features. MacMillan (1987) concluded, "Within-category feature frequencies are retained only if they are useful in distinguishing that category. If feature frequencies are irrelevant in distinguishing category members from nonmembers, there is no need to retain feature frequency information that is sorted by category" (p. 38). MacMillan's results disconfirmed theories such as Estes's (1959) feature-frequency model and Medin and Schaffer's (1978) context model, which store veridical frequency counts; she showed instead that her results, that frequency counts depend on the validity of the cue, were well fit by our adaptive network model with the LMS learning rule.

Extensions of the Model

We have already noted that the network model we tested has a simple architecture and identifies stimulus patterns (such as one or more symptoms) in a simple, direct manner. We are currently investigating several modifications or amendments of the simple network theory by way of computer simulations. One modification assumes that of each trial one or more background stimuli are presented independently of whatever symptom pattern the experimenter is presenting. This background cue serves the same role as the experimental context or the "conditioning apparatus" in the studies by Rescorla and Wagner (1972). Our analyses indicate that the addition of a context cue does not improve the model's ability to predict our data.[5] For example, for Experiment 1 the model with the context cue predicts probability ratings that deviate about twice as far from the data as do the predictions of the model without the context cue. Moreover, the context-cue model predicts in Experiment 1 that symptom s_2 would become predominately associated with the rare disease, a prediction not supported by data. Such predictions arise because the omnipresent context cue would become strongly associated with the common disease. Thus the predicted association between the symptoms and the common disease would be attenuated by the competitive nature of the LMS algorithm. Compensating for this attenuated association with the common disease, the context-cue model expects the symptoms to become more strongly associated with the rare disease. But this pattern of changes caused the context-cue model to predict the data more poorly than did the simpler model.

Another theoretical variation we are investigating assumes that sensory units themselves are interconnected by links with adjustable weights. Whenever two sensory units are on together or off together, the link between them will be strengthened. Such a network of intersensory connections can learn symptom–symptom correlations across patterns as well as symptom–disease correlations (as our current network does). Such amendments will presumably be required for a one-layer model to handle results by Medin, Altom, Edelson, and Freko (1982) and others who have demonstrated subjects' sensitivity to such symptom correlations.

In conclusion, we have tentatively accepted the LMS error criterion as confirmed by the results of Experiments 1, 2, and 3. We are currently conducting further tests of implications of the LMS rule in the symptom–disease learning paradigm.

[5] The addition of a constant context cue to the network model has an effect similar to that of adding a constant to the linear-regression model.

We are encouraged not only that the LMS rule of connectionist theories that fits human learning data links with the Rescorla–Wagner (1972) model of conditioning but also that it implies the phenomenon of base-rate neglect, which has proved to be a robust phenomenon in the literature of judgment and decision. Such theoretical connections across disparate research areas are especially encouraging to the goals of cognitive psychology.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147–169.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

Baker, T. W., & Mackintosh, N. S. (1977). Excitatory and inhibitory conditioning following uncorrelated presentations of CS and UCS. *Animal Learning and Behavior, 5,* 315–319.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologia, 44,* 211–233.

Borgida, E., & Brekke, N. (1981). The base-rate fallacy in attribution and prediction. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 63–95). Hillsdale, NJ: Erlbaum.

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.

Dickinson, A., & Shanks, D. (1985). Animal conditioning and human causality judgment. In L. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory.* Hillsdale, NJ: Erlbaum.

Ebbinghaus, H. (1885). *Über das Gedächtnis* [On memory]. Leipzig, GDR: Duncker und Humbolt.

Estes, W. K. (1959). Component and pattern models with Markovian interpretation. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 9–52). Stanford, CA: Stanford University Press.

Estes, W. K. (1985). Some common aspects of models for learning and memory in lower animals and man. In L. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 151–166). Hillsdale, NJ: Erlbaum.

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

Feigenbaum, E. A. (1959). *An information processing theory of verbal behavior.* Santa Monica, CA: Rand Corp.

Feigenbaum, E. A., & Simon, H. A. (1961). Forgetting in an associative memory. *Proceedings of the ACM National Conference, 16,* 202–205.

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.

Gluck, M. A. (1984). [Subjects trained to discriminate different levels completely transfer these values when they later estimate the reverse probabilities]. Unpublished raw data.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

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.

Johnson, R. A., & Wichern, D. W. (1982). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice-Hall.

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.

Kassin, S. M. (1979). Base rates and prediction: The role of sample size. *Personality and Social Psychology Bulletin, 5,* 210–213.

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

Le Cun, Y. (1985). Une procedure d'apprentissage pour reseau a seuil assymetrique [A procedure for learning an asymmetric threshold network]. *Proceedings of Cognitiva 85, Paris,* 599–604.

Mackintosh, N. J., (1983). *Conditioning and associative learning.* Oxford, England: Oxford University Press.

Mackintosh, N. J., & Honig, W. K. (1969). *Fundamental issues in associative learning.* Halifax, Nova Scotia, Canada: Dalhousie University Press.

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

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.

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. (1984). Learning of ill-defined categories by monkeys. *Canadian Journal of Psychology, 38,* 285–303.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117,* 68–85.

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

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

Neeley, J. H. (1982). The role of expectancy in probability learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 599–607.

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

Parker, D. (1985). Learning logic (Report No. 47). Cambridge, Massachusetts Institute of Technology, Center for Computational Research in Economics and Management Science.

Parker, D. (1986). A comparison of algorithms for neuron-like cells. In J. Denker (Ed.), *Proceedings of the Neural Networks for Computing Conference.* New York: American Institute of Physics.

Pavlov, I. (1927). *Conditioned reflexes.* London: Oxford University Press.

Prokasy, W. F. (1965). Classical eyelid conditioning: Experimental operations, task demands, and response shaping. In W. F. Prokasy (Ed.), *Classical conditioning* (pp. 208–225). New York: Appleton-Century-Crofts.

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

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1–5.

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.

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

Rudy, J. W. (1974). Stimulus selection in animal conditioning and paired-associate learning: Variations in the associative process. *Journal of Verbal Learning and Verbal Behavior, 13,* 282–296.

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

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: Bradford Books/MIT Press.

Schank, R. C. (1982). *Dynamic memory.* Cambridge, England: Cambridge University Press.

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. (1956). *Behavior theory and conditioning.* New Haven, CT: Yale University Press.

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* (pp. 444–459). Cambridge, MA: Bradford Books/MIT Press.

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

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). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, England: Cambridge University Press.

Wagner, A. R. (1969). Stimulus selection and a modified continuity theory. In G. Bower & J. Spence (Eds.), *The psychology of learning and motivation* (Vol. 3, pp. 1–41). New York: Academic Press.

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.

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

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology, 86,* 837–845.

Appendix A

To find asymptotic values for the weights in the adaptive networks used here, we derive an equation for the expected value of the weights at time t and then let t become large. This derivation was adapted from the work of Parker (1986) and Stone (1986). The basic least mean squares (LMS) learning rule says that the change in weights, Δw_{ij}, from input node i to output node j is governed by

$$\Delta w_{ij} = \beta\left(d_j - \sum_{k=1}^{n} w_{kj}a_k\right)a_i, \qquad \text{(A-1)}$$

where a_i is the activation on input node i, d_j is the desired output for node j, and the summation is over the n input nodes. Because the weights are updated at discrete trial intervals $t = 1, 2, \ldots$, we index Equation A-1 to make this trial dependence explicit, where

$$\Delta w_{ij}(t) = \beta\left[d_j(t) - \sum_{k=1}^{n} w_{kj}(t)a_k(t)\right]a_i(t). \qquad \text{(A-2)}$$

We can recast Equation A-2 in matrix notation by using the column vectors $W_j(t)$ and $A(t)$, where

$$W_j(t) = \begin{bmatrix} w_{1j}(t) \\ \vdots \\ w_{ij}(t) \\ \vdots \\ w_{nj}(t) \end{bmatrix} \quad \text{and} \quad A(t) = \begin{bmatrix} a_1(t) \\ \vdots \\ a_i(t) \\ \vdots \\ a_n(t) \end{bmatrix}$$

We can now rewrite Equation A-2 in vector notation as

$$\Delta W_j(t) = \beta[d_j(t) - W_j(t)^T A(t)]A(t), \qquad \text{(A-3)}$$

where $W_j(t)^T$ is the transpose of $W_j(t)$. Equivalently,

$$\begin{aligned} W_j(t+1) &= W_j(t) + \beta[d_j(t) - W_j(t)^T A(t)]A(t) \\ &= W_j(t) + \beta d_j(t)A(t) - \beta[W_j(t)^T A(t)]A(t). \end{aligned}$$

Note that $W_j(t)^T A(t)$ is a scalar multiplier and can be reexpressed as

$$W_j(t)^T A(t) = \sum_{k=1}^{n} w_{kj}(t)a_k(t) = A(t)^T W_j(t).$$

With this equality, we can regroup the terms in Equation A-3 so that

$$W_j(t+1) = \beta d_j(t)A(t) + [I - \beta A(t)A(t)^T]W_j(t), \qquad \text{(A-4)}$$

where I is the identity matrix and $A(t)A(t)^T$ is the symptom covariance matrix,

$$A(t)A(t)^T = \begin{bmatrix} a_1(t)^2 & \cdots & a_1(t)a_n(t) \\ \vdots & \ddots & \vdots \\ a_n(t)a_1(t) & \cdots & a_n(t)^2 \end{bmatrix}$$

Equation A-4 is recursive in that the expression for $W_j(t+1)$ depends on $W_j(t)$ and also $A(t)$ and $d_j(t)$. For stationary stochastic inputs, the expected matrices $E[A(t)]$ and $E[d_j(t)]$ are independent of t, so we may rewrite $A(t)$ as A and $d_j(t)$ as d_j, and $E[A(t)A(t)^T W_j(t)]$ becomes $E[AA^T]E[W_j(t)]$. Hence

$$E[W_j(t+1)] = \beta E[d_j A] + \{I - \beta E[AA^T]\}E[W_j(t)]. \qquad \text{(A-5)}$$

Because the inputs are stochastically stationary, it can be shown that as long as β is small enough that the magnitudes of the eigenvalues of $\{I - \beta E[AA^T]\}$ are less than 1, the weights will converge and $\lim_{t\to\infty} E[W_j(t+1)] = \lim_{t\to\infty} E[W_j(t)] = E[W_j]$ (Parker, 1986). Hence

$$E[W_j] = \beta E[d_j A] + \{I - \beta E[AA^T]\}E[W_j]. \qquad \text{(A-6)}$$

If all the rows (or, equivalently, columns) of $E[AA^T]$ are linearly independent, then $E[AA^T]$ is invertible, and a unique LMS solution exists for the weights. If $E[AA^T]$ is not invertible, then see below. Equation A-6 reduces to

$$E[W_j] = E[AA^T]^{-1}E[d_j A], \qquad \text{(A-7)}$$

which is the optimal LMS solution to the learning problem. With sufficiently small β, the LMS rule will converge to this solution regardless of the initial configuration of the weights. If $E[AA^T]$ is not invertible (i.e., the determinant is zero), then there exists more than one set of weights that will give an LMS solution (see Appendix B). In this latter case, the convergence of the network does depend on the initial weights.

From Equation A-7, we can derive a closed-form expression for the asymptotic values of the weights for categorization problems in which the likelihoods of the features (symptoms) within a category (disease) are independent of each other. This expression will be in terms of the overall probabilities of the categories (i.e., the base-rate frequencies) and the probabilities of the features given the categories.

Because this derivation and the expressions are quite cumbersome for four symptoms, we illustrate the method by deriving an expression for the weights in a simpler arrangement that uses only two symptoms, s_1 and s_2, and two disease categories, C_1 and C_2. As noted earlier, the asymptomatic weights in a network for a forced choice between two mutually exclusive alternatives (using $+1/-1$ as feedback) can be derived from a network in which there are two output nodes, one for each choice. Each of the output nodes in such a network corresponds to one of the two categories, receiving a $+1$ reinforcement in the presence of that category and a 0 reinforcement in its absence. If these two output nodes are C_1 and C_2, then w_i, the asymptotic weight of the association between s_i and the output node in the single-output note network, equals $w_{i1} - w_{i2}$, the difference between the association strengths between s_i and the two output nodes in the two-output node model. Because of this, we simply derive an expression for the weights into one of the output nodes (in the two-output node network), for example, C_1, assuming that it receives feedback of 1 when C_1 occurs and 0 when C_2 occurs. In this case, the relevant terms of Equation A-7 become

$$\begin{aligned} E[AA^T]^{-1} &= \begin{bmatrix} P(s_1) & P(s_1 \,\&\, s_2) \\ P(s_1 \,\&\, s_2) & P(s_2) \end{bmatrix}^{-1} \\ &= \frac{1}{P(s_1)P(s_2) - P(s_1 \,\&\, s_2)^2}\begin{bmatrix} P(s_2) & -P(s_1 \,\&\, s_2) \\ -P(s_1 \,\&\, s_2) & P(s_1) \end{bmatrix}. \end{aligned} \qquad \text{(A-8)}$$

For C_1,

$$E[d_1 A] = \begin{bmatrix} P(s_1 \,\&\, C_1) \\ P(s_2 \,\&\, C_1) \end{bmatrix}. \qquad \text{(A-9)}$$

Thus, by combining Equations A-8 and A-9, we obtain the weights of the cues toward C_1 of

$$E[W_{j1}] = E[AA^T]E[d_1A]$$

$$= \frac{\begin{bmatrix} P(s_1 \,\&\, C_1)P(s_2) - P(s_1 \,\&\, C_1)P(s_1 \,\&\, s_2) \\ -P(s_2 \,\&\, C_1)P(s_1 \,\&\, s_2) + P(s_2 \,\&\, C_1)P(s_1) \end{bmatrix}}{P(s_1)P(s_2) - P(s_1 \,\&\, s_2)^2} \quad \text{(A-10)}$$

Considering the top term of Equation A-10, the association between s_1 and C_1, we obtain

$$E[w_{11}] = \frac{P(s_1 \,\&\, C_1)P(s_2) - P(s_2 \,\&\, C_1)P(s_1 \,\&\, s_2)}{P(s_1)P(s_2) - P(s_1 \,\&\, s_2)^2}$$

$$= \frac{\left[\frac{P(s_1 \,\&\, C_1)P(s_2)}{P(s_1)P(s_2)}\right] - \left[\frac{P(s_2 \,\&\, C_1)P(s_1 \,\&\, s_2)}{P(s_1)P(s_2)}\right]}{1 - \frac{P(s_1 \,\&\, s_2)^2}{P(s_1)P(s_2)}}$$

$$= \frac{P(C_1 \mid s_1) - P(C_1 \mid s_2)P(s_2 \mid s_1)}{1 - P(s_1 \mid s_2)P(s_2 \mid s_1)} \quad \text{(A-11)}$$

Equation A-11 shows that the expected value of the weight between symptom s_1 and disease category C_1 is proportional to the evidential weight of s_1 for C_1, for example, $P(C_1 \mid s_1)$, minus the evidential weight of the other symptom, s_2, for that category weighted to reflect the degree to which s_2 co-occurs with s_1. The asymptotic weights are also inversely proportional to one minus the contingency between the two symptoms, for example, $P(s_1 \mid s_2)P(s_2 \mid s_1)$; but, because the denominator is common to all weights in the network, this term does not influence comparisons among weights. If one increases the expression to include three and four symptoms, one will see analogously that the overall weights are determined by the conditional probabilities of C_1 given s_1 minus the conditional probabilities of the other symptoms, weighted by their independent degrees of co-occurrence with s_1.

To numerically calculate the actual asymptotic values for the four symptoms from Experiment 1, go back to Equation A-7. One can express $E[AA^T]$ and $E[dA]$ in terms of the overall frequencies of the disease categories, $P(R)$ and $P(C)$, and the likelihoods of the four symptoms given the diseases. Because $a_i = 1$ if s_i is present and 0 if s_i is absent and because $P(R) + P(C) = 1$, one can derive that $E[a_ia_j] = P(s_i \,\&\, s_j \mid R)P(R) + P(s_i \,\&\, s_j \mid C)P(C)$ if i is not equal to j and that $E[a_ia_j] = E[a_i^2] = P(s_i \mid R)P(R) + P(s_i \mid C)P(C) = P(s_i)$ if $i = j$. We wish to obtain the symptom covariance matrix, $E[AA^T]$, which, as noted before, is

$$E[AA^T] = \begin{bmatrix} E[a_1^2] & \cdots & E[a_4a_1] \\ \vdots & \ddots & \vdots \\ E[a_1a_4] & \cdots & E[a_4^2] \end{bmatrix}.$$

Using the relations noted already plus the probabilities from Experiment 1, one obtains

$$E[AA^T] = \begin{bmatrix} .347 & .121 & .121 & .139 \\ .123 & .375 & .139 & .179 \\ .121 & .139 & .433 & .225 \\ .139 & .179 & .225 & .578 \end{bmatrix}.$$

Similarly, one can calculate the correlations between the training signals and the symptoms, $E[dA]$. Because $d = 1$ if Disease R is the correct diagnosis and -1 if Disease C is correct, one will have $E[da_i] = P(s_i \mid R)P(R) - P(s_i \mid C)P(C)$. Using the probabilities from Experiment 1, one has

$$E[dA] = \begin{bmatrix} P(s_1 \mid \mathrm{R})P(\mathrm{R}) - P(s_1 \mid \mathrm{C})P(\mathrm{C}) \\ P(s_2 \mid \mathrm{R})P(\mathrm{R}) - P(s_2 \mid \mathrm{C})P(\mathrm{C}) \\ P(s_3 \mid \mathrm{R})P(\mathrm{R}) - P(s_3 \mid \mathrm{C})P(\mathrm{C}) \\ P(s_4 \mid \mathrm{R})P(\mathrm{R}) - P(s_4 \mid \mathrm{C})P(\mathrm{C}) \end{bmatrix} = \begin{bmatrix} 0 \\ -.144 \\ -.260 \\ -.462 \end{bmatrix}.$$

Substituting these into Equation A-7, one obtains the asymptotic values of the weights for the one-output node network:

$$W = \begin{bmatrix} .347 & .121 & .121 & .139 \\ .121 & .357 & .139 & .179 \\ .121 & .139 & .433 & .225 \\ .139 & .179 & .225 & .578 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -.144 \\ -.260 \\ -.462 \end{bmatrix} \approx \begin{bmatrix} .430 \\ -.043 \\ -.306 \\ -.771 \end{bmatrix},$$

and these are plotted in Figure 3. The asymptotic weights for Experiment 2 can be derived in a similar manner.

Appendix B

To derive the asymptotic values of the weights for Experiment 3, one proceeds as in Appendix A, calculating the expected symptom covariance matrix:

$$E[AA^T] = \begin{bmatrix} .3 & .105 & .105 & .12 & 0 & .195 & .195 & .18 \\ .105 & .325 & .12 & .155 & .22 & 0 & .205 & .17 \\ .105 & .12 & .375 & .195 & .27 & .255 & 0 & .18 \\ .12 & .155 & .195 & .5 & .38 & .345 & .305 & 0 \\ 0 & .22 & .27 & .38 & .7 & .48 & .43 & .32 \\ .195 & 0 & .255 & .345 & .48 & .675 & .42 & .33 \\ .195 & .205 & 0 & .305 & .43 & .42 & .625 & .32 \\ .18 & .17 & .18 & 0 & .32 & .33 & .32 & .5 \end{bmatrix}.$$

One then calculates the covariance vector between the training signals and the symptoms:

$$E[dA] = \begin{bmatrix} 0 \\ -0.125 \\ -0.225 \\ -0.4 \\ -0.5 \\ -0.375 \\ -0.275 \\ -0.1 \end{bmatrix}.$$

At this point, apply Equation A-7 to calculate the asymptotic weights. In contrast to the analyses of Experiments 1 and 2, however, the stimulus environment for Experiment 3 underdetermines the asymptotic weights. In this situation, the asymptotic values depend on the initial values of the weights. (This situation should be recognizable to those familiar with the model's analysis of the Kamin, 1969, blocking

experiment.) In a situation such as this, $E[AA^T]$ is not invertible. If, however, one assumes that the initial weights are all zero, then it can be shown (Parker, 1986) that instead of the inverse, $E[AA^T]^{-1}$, it is permissible to use the pseudoinverse, $E[AA^T]^+$, where

$$E[AA^T]^+ = \begin{bmatrix} 1.54 & .268 & .312 & .366 & -1.13 & .147 & .103 & .049 \\ .268 & 1.42 & .262 & .262 & .100 & -1.06 & .106 & .139 \\ .312 & .262 & 1.298 & 1.298 & .004 & .055 & -.981 & .171 \\ .366 & .229 & .146 & .146 & -.161 & -.024 & .060 & -.932 \\ -1.126 & .100 & .004 & .004 & 1.06 & -.166 & -.070 & .095 \\ .147 & -1.06 & .055 & .055 & -.166 & 1.04 & -.073 & .005 \\ .103 & .106 & -.981 & -.981 & -.070 & -.073 & 1.02 & -.026 \\ .049 & .139 & .171 & .171 & .095 & .005 & -.026 & 1.08 \end{bmatrix}$$

When the stimulus environment underdetermines the asymptotic weights, there is an infinite set of solutions that minimize the expected mean squared error; under these circumstances the pseudoinverse provides a solution with the smallest magnitude of weights, as measured by the sum of the squares of the weights. When one applies Equation A-7 with the pseudoinverse instead of the inverse, the asymptotic values of the weights are

$$W = \begin{bmatrix} .224 \\ -.026 \\ -.152 \\ -.350 \\ -.408 \\ -.156 \\ -.030 \\ .166 \end{bmatrix}$$

and these points are plotted in Figure 7. Simulation results provide further confirmation that these are, in fact, the asymptotes reached when the network is started with all zero weights.

To explore the network model's sensitivity to the initial configuration of weights, we began by training the network to asymptote by using reversal category feedback; that is, we reinforced rare-disease trials with -1 instead of $+1$ and common-disease trials with $+1$ instead of -1. This resulted in weights that are just the negation of the asymptotes given earlier. Using these weights as an initial configuration, we then switched back to the normal training procedure. The network subsequently asymptoted at the original weights already given. Because the asymptotic weights from the first phase of training were just the reverse of what is actually required for the second phase of training, we take these results to suggest that the weights originally derived are quite robust to variations in the initial weights.

Received November 19, 1986
Revision received November 25, 1987
Accepted March 23, 1988 ■