From Learning Theory to Connectionist Theory

Essays in Honor of William K. Estes

Volume 1

Edited by
Alice F. Healy
Stephen M. Kosslyn
Richard M. Shiffrin

9 Stimulus Sampling and Distributed Representations in Adaptive Network Theories of Learning

Mark A. Gluck
Rutgers University
Stanford University

Current adaptive network, or "connectionist," theories of human learning are reminiscent of statistical learning theories of the 1950s and early 1960s, the most influential of which was *Stimulus Sampling Theory*, developed by W. K. Estes and colleagues (Atkinson & Estes, 1963; Estes, 1959). Both Stimulus Sampling Theory and adaptive network theory are general classes of learning theories—formal frameworks within which theorists search for a small number of concepts and principles that will illuminate a wide variety of psychological phenomena when applied in varying combinations. To the extent that adaptive networks represent cumulative progress in theory development, we should expect them to incorporate the strengths of Stimulus Sampling Theory but overcome the problems that limited these earlier approaches to modeling associative learning.

This chapter reviews Stimulus Sampling Theory (SST), noting some of its strengths and weaknesses, and compares it to a recent network model of human learning (Gluck & Bower, 1986; 1988a). We will see that the network model's learning rule for updating associative weights represents a significant advance over Stimulus Sampling Theory's more rudimentary learning procedure. In contrast, Stimulus Sampling Theory's stochastic scheme for representing stimuli as distributed patterns of activity can overcome some limitations of network theories that identify stimulus cues with single active input nodes. This leads us to consider a distributed network model that embodies the processing assumptions of our earlier network model but employs stimulus-representation assumptions adopted from Stimulus Sampling Theory. In this distributed network, stimulus cues are represented by the stochastic activation of overlapping populations of stimulus elements (input nodes). Rather than replacing the two previous learning theories, this distributed network combines the best established concepts of the

earlier theories and reduces to each of them as special cases in those training situations where the previous models have been most successful.

STIMULUS SAMPLING THEORY

Stimulus Sampling Theory treats learning as a stochastic process in which stimuli are represented as populations of independent variables, called *stimulus elements* (for reviews, see Bower & Hilgard, 1981; Neimark & Estes, 1967). On any individual experimental trial, only a subset of these elements is presumed to be sampled by the subject. Each element in the set is assumed to be completely associated with one of the possible responses available to the subject. Traditionally, two different, but often functionally equivalent modeling schemes have been employed to describe how stimulus elements are sampled. One scheme supposes that each stimulus element has a certain independent probability of being sampled whereas the other scheme assumes that a fixed number of stimuli are taken from the total population on each trial. Once a subset of the population has been sampled, choice behavior is determined by the proportion of sampled elements associated with each response. For example, if 75% of the sampled elements are associated with response R_1 and 25% with response R_2, the model predicts that the subject will respond R_1 with probability .75. Reinforcement occurs through the total conditioning of all sampled elements, each of which becomes associated with the reinforced outcome.

Probability Learning

Many early applications of Stimulus Sampling Theory were concerned with probability learning experiments where subjects were trained to predict which of several randomly chosen outcomes would occur. The most basic probability learning situation involves two possible outcomes, which we refer to as E_1 and E_2. At the beginning of each trial, subjects give one of two possible responses: R_1 if they expect E_1, and R_2 if they expect E_2. The stimulus conditions in this experiment are represented by a single population of stimulus elements. If p_n denotes the proportion of elements connected to R_1 at the beginning of trial n, then p_n is both the probability that a randomly selected element will be connected to R_1 and the probability that the subject will respond R_1 on that trial. If θ is the probability that an individual element is sampled, the theory implies a learning equation described by changes in p_n where

$$p_{n+1} = \begin{cases} (1 - \theta)p_n + \theta & \text{if } R_1 \text{ is reinforced} \\ (1 - \theta)p_n & \text{if } R_2 \text{ is reinforced.} \end{cases} \quad (1)$$

If the reinforcing event, E_1, occurs with constant probability π, we can rewrite Equation 1 as

$$p_{n+1} = \pi[(1 - \theta)p_n + \theta] + (1 - \pi)(1 - \theta)p_n$$
$$= (1 - \theta)p_n + \theta\pi \quad (2)$$

Following extended training, Equation 2 predicts that p_n, the probability of responding R_1, will come to match π, the objective probability of E_1 occurring. One way to see this is to compute the long-term average proportion of E_1 events (π) and show that this is equivalent to $\left(\frac{1}{T}\sum_{n=1}^{T} p_n\right)$, the long-term average probability of responding R_1 (see Hilgard & Bower, 1975, p. 386). Alternatively, we can rewrite Equation 2 in terms of the expected change in p, Δp, from trial n to trial $n + 1$,

$$\Delta p = \theta(\pi - p_n). \quad (3)$$

Equations 1 and 3 characterize a "linear model" of learning, expressing how p_n, the probability of responding R_1, changes as a linear function of p_{n-1}. From Equation 3 we see that the system will stabilize (i.e., the expected change in p,$\Delta p = 0$) when $p_n = \pi$. This implies that p_n, the proportion of R_1 responses, will come to match π, the average proportion of E_1 events. This appears to be a non-optimal strategy because "probability matching" generally yields a lower expected proportion of correct responses compared to a "probability maximizing" strategy in which a subject always chooses the most likely outcome. This is most easily seen with reference to a two-choice task. A correct response occurs whenever the subject responds R_1 on an E_1 trial (which will occur with probability π^2), or when the subject responds R_2 on an E_2 trial (which will occur with probability $(1 - \pi)^2$). Thus, probability matching yields an expected proportion correct of $\pi^2 + (1 - \pi)^2$. Note that this is always less than or equal to the proportion correct expected from adopting a probability maximizing strategy. For example, if $\pi = .8$, a probability matching strategy results in a long-run average of .68 correct whereas a probability maximizing strategy (choosing R_1) yields a long-run average of .8 correct.

The probability matching prediction of Stimulus Sampling Theory has been tested in a wide variety of training situations and, for the most part, these predictions have been confirmed. For example, Suppes and Atkinson (1960, p. 196) report an experiment in which $\pi = .60$ and the observed proportion of R_1 responses was .596 (averaged for 30 subjects over the last 100 out of 240 trials). The model also makes fairly accurate predictions regarding the shape of learning curves under a variety of reinforcement schedules (Estes, 1964). Other tests of Stimulus Sampling Theory have demonstrated its ability to predict sequential statistics that describe the extent to which a subject's response on trial $n + 1$ is influenced by his responses and/or the reinforcing events on trial n (Atkinson, Bower, & Crothers, 1965; Suppes & Atkinson, 1960). The theory has also been applied, with considerable success, to such diverse phenomena as spontaneous recovery and forgetting (Estes, 1955), reaction time distributions (Bush & Mosteller, 1955), and recognition memory (Bower, 1972).

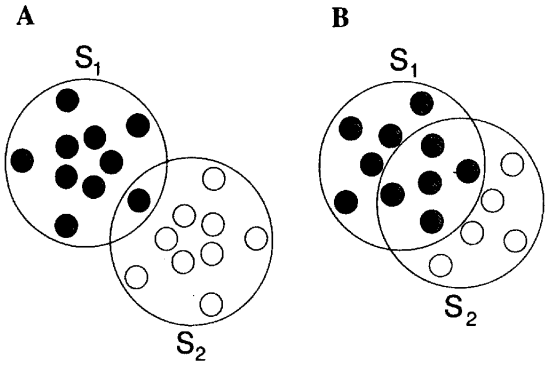

FIG. 9.1. Generalization through shared elements for two stimuli S_1 and S_2 that are: (A) slightly similar (small overlap) or (B) very similar (large overlap).

Stimulus Generalization and Discrimination Learning

In Stimulus Sampling Theory, *stimulus generalization* between distinct stimuli is conceived to arise from common elements shared by both stimuli, an approach drawn from Thorndike's (1898) "connectionism" theory of Stimulus-Response formation. For example, the response associated with stimulus S_1 in Fig. 9.1 will generalize to stimulus S_2 to the extent that they share common stimulus elements. It has long been known that if a stimulus such as a high frequency tone, S_1, is associated with some significant event (R_1), this conditioning will generalize to other similar stimuli such as a low tone (S_2). Within Stimulus Sampling Theory the generalization of the $S_1 \rightarrow R_1$ association to an $S_2 \rightarrow R_1$ association is predicted to be in direct proportion to the amount of overlap between the S_1 and S_2 stimulus pools. This common-elements approach to stimulus generalization has been successfully applied to a range of generalization phenomena (Atkinson & Estes, 1963; LaBerge, 1961).

The application of this approach to *discrimination learning* has, however, been problematic. Consider the case in which stimulus S_1 is paired with response R_1 and stimulus S_2 is paired with response R_2. Stimulus Sampling Theory's learning rule expects the distinctive elements in S_1 and S_2 to become totally conditioned to R_1 and R_2, respectively. The common elements, however, will be conditioned to both responses in weighted proportion to the R_1 and R_2 reinforcement frequencies. Responding to stimuli will therefore be controlled by the correctly conditioned distinct elements as well as the "mixed conditioned" common elements. Thus, the model predicts, incorrectly, that perfect discrimination should never occur. People and animals can, of course, be trained to respond differentially to distinctive stimuli, even when the stimuli are quite similar.

It might seem simple enough to propose a mechanism for discrimination learning whereby the subject learns to "adapt out" the shared common cues and

attend only to the unique cues. Several theorists have proposed "attentional" extensions to Stimulus Sampling Theory. One approach is to postulate additional mechanisms that render common elements ineffective during the course of discrimination learning (Bush & Mosteller, 1955; Lovejoy, 1968; Restle, 1957; Sutherland & Mackintosh, 1971). Other approaches include the addition of specialized observing responses (Atkinson, 1958; Levine, 1970) or modified decision processes (Laberge, 1962). One problem with these approaches is that by nullifying the effect of the shared common elements, they predict that these common elements will have no influence on transfer tasks. Several studies have shown that even with complete discrimination training, the common elements can still exert a strong influence on subsequent transfer tasks when subjects are asked to classify novel combinations of cues (Binder & Feldman, 1960; Binder & Taylor, 1969; Flagg & Medin, 1973; Robbins, 1970).

Estes (1959) considered still another approach in which the stimulus situation on a given trial is viewed as a unique pattern rather than as a collection of component cues. This pattern model has a desirable property that the component-cue model lacks: the ability to predict perfect discrimination between two stimulus patterns in the presence of common elements. Mitigating this advantage, however, is the failure of the pattern model to provide an adequate account of stimulus generalization. A natural combination of the component and pattern theories is the mixed model first proposed by Estes and Hopkins (1961) and later developed quantitatively by Atkinson and Estes (1963). According to this model, associations are formed during discrimination training, between the component cues and the responses as well as between the pattern cues and the responses. Once the pattern cues are learned, they are presumed to dominate in discrimination tasks. In generalization tasks, however, the component cues mediate responding. Whereas the mixed model has had some success in resolving the overlap problem, the interactions between the component and pattern processes have not been completely evaluated for the full range of discrimination and generalization tasks. In summary, an entirely satisfactory resolution to the "overlap problem" has not been developed that successfully reconciles stimulus generalization and discrimination learning (Bower & Hilgard, 1981; Medin, 1976).

Binder and Estes (1966): A Stimulus-Sampling-Theory Interpretation

To better appreciate the subtleties involved in trying to resolve the overlap problem in discrimination and generalization, we now consider in detail a study by Binder and Estes (1966). This study illustrates an additional problem with Stimulus Sampling Theory; its inability to account for a phenomena that Binder and Estes termed the "relative novelty" effect. We describe this study (and subsequent extensions and elaborations by other investigators) and then use these data as a test base to compare Stimulus Sampling and adaptive network interpretations of discrimination learning.

Binder and Estes (1966) conducted systematic studies of the effects of category frequency on learning, following a line of research begun by Binder and Feldman (1960). Subjects were trained to classify patterns composed of several simple component cues. Following this training, they classified novel combinations of the component cues. Stimulus patterns *ab* and *ac* were reinforced with responses R_1 and R_2, respectively. The critical manipulation was the unbalanced presentation frequencies of the different reinforcements; $ab \rightarrow R_1$ trials occurred three times as often as $ac \rightarrow R_2$ trials.

A Stimulus-Sampling-Theory interpretation of this experiment posits three populations (pools) of elements corresponding to each of the three cues. Presentation of the *ab* pattern activates elements in both the *a* pool and the *b* pool. With the unbalanced presentation frequencies of the two reinforcements, Stimulus Sampling Theory predicts that after extended training, all stimulus elements from the *b* pool will be associated with R_1, the more common outcome, and all elements from the *c* pool will be associated with R_2, the less common outcome. The *a* pool, however, will contain some elements associated with R_1 and others associated with R_2. Because R_1 was presented three times as often as R_2, Stimulus Sampling Theory predicts that 75% of the elements in the *a* pool will be conditioned to R_1, and 25% will be conditioned to R_2. Therefore, Stimulus Sampling Theory expects that the presentation of symptom *a* alone should, on average, activate a stimulus sample with the majority of elements predicting the common category. This result accords with data from studies by Binder and Feldman (1960) who used a 2:1 ratio of $ab \rightarrow R_1$ to $ac \rightarrow R_2$ presentations and observed response proportions for the shared common cue (*a*) of .65 and .29 for R_1 and R_2, respectively (compared to predicted values of .67 and .33). With a 4:1 ratio of presentations frequencies, they observed response proportions for the common cue of .76 and .20 (compared to predicted values of .80 and .20).

As noted earlier, Stimulus Sampling Theory is unable to account for people's ability to discriminate between similar stimuli that activate overlapping populations of hypothetical elements. The same problem exists when the common elements are explicit, as in the Binder and Feldman (1960) and Binder and Estes (1966) studies. As Binder and Estes noted, as long as subjects are randomly sampling elements from the *a* pool, it is possible that *ab* patterns will be incorrectly classified with the rare outcome and *ac* patterns incorrectly classified with the common outcome (p. 3). Not surprisingly, subjects learned to master perfectly the $ab \rightarrow R_1/ac \rightarrow R_2$ discrimination with relative ease. Because of this shortcoming of the Stimulus Sampling model, Binder and Estes suggested that it might be necessary to augment Stimulus Sampling Theory with a mechanism by which subjects could learn to "respond selectively to cues which are reliable predictors of reinforcing events . . . and to ignore or 'adapt to' common cues . . . which are not uniformly correlated with reinforcement" (p. 4). As described earlier, several schemes for doing just this were proposed in the literature. However, the perfect discrimination attained by subjects in this task would seem to suggest that subjects had learned to "adapt to" or ignore the common *a*

cue. This is inconsistent with the previously described transfer effect wherein subject's transfer classification of the *a* cue indicates that they have clearly learned that *a* predicts R_1, the more common reinforcement. Thus, the use of explicit common elements (cues), as in the Binder and Feldman (1960) and Binder and Estes (1966) studies makes clear the difficulty in reconciling discrimination and generalization behaviors.

The Relative Novelty Effect

Stimulus Sampling Theory also fails to account for another aspect of Binder and Estes' data. During the transfer task, subjects were given the the novel feature combination (*bc*). Stimulus Sampling Theory expects that early in training *bc* should be more associated with R_1, the more common outcome, because there will be more *b* elements associated with R_1 than *c* elements associated with R_2. Once learning is complete, however, Stimulus Sampling Theory predicts that *bc* should, on average, activate an equal number of oppositely associated elements, predicting that *bc* should be equally associated with the two reinforcements. In summary, Stimulus Sampling Theory predicts that *bc* will be associated either with the more frequent reinforcement (R_1) or with both reinforcements equally. Surprisingly, subjects were more likely to classify the *bc* pattern with R_2, the *less* frequent outcome. Binder and Estes called this the "relative-novelty" effect because the probability of a stimulus component controlling choice behavior (*c* in this case) appears to be inversely related to its presentation frequency during training. Subsequent replications and extensions of this result have been presented by Binder and Taylor (1969), Medin and Robbins (1971), and Medin and Edelson (1988). Heretofore, no satisfactory explanation has been offered for it.

GENERALIZATION AND DISCRIMINATION IN ADAPTIVE NETWORKS

Stimulus Sampling Theory was strongly motivated by the principle that

> . . . we can hope to understand the processes that guide adult human behavior only within a rather broad framework in which they can be meaningfully related both to the more primitive or elementary processes from which they develop during the life of the individual and to those of lower organism. . . . Typically, evolution works through endless variations on a limited repertory of themes . . . as a consequence, clues to understanding complex processes of human cognition sometimes come from studying simpler forms. (Estes, 1982, p. 315)

Subsequent developments in learning theory all but abandoned this unified approach to understanding human and infrahuman learning. Animal research continued to be primarily concerned with elementary associative processes whereas, by the mid-1960s, human learning began to be characterized in terms of informa-

tion processing and hypothesis testing, concepts borrowed from artificial intelligence and computer science. The recent emergence of "parallel distributed processing" models based on "connectionist" networks, however, presents an alternative to the rule-based symbolic models of the 1970s and early 1980s. Like the earlier statistical learning theories, these network models embody the assumption that many complex human abilities can best be understood as emerging from configurations of elementary associative processes.

In recent papers, we have used adaptive networks to explore the relationship between human learning and the elementary associative learning processes that can be studied in simpler organisms. We began by studying a simple adaptive network model of human learning that extends Rescorla and Wagner's (1972) description of classical conditioning to human classification learning (Gluck & Bower, 1986, 1988a, 1988b; Gluck, Bower, & Hee, 1989). The learning rule is the same as the least mean squares (LMS) learning rule for training one-layer networks, first proposed by Widrow and Hoff (1960). The model has been fit to data from experiments on probabilistic classification learning with multiple cues. Although this simple model can be applied only to a restricted range of experimental circumstances, it has shown a surprising accuracy in predicting human behavior within that range, including data on people's choice proportions during learning, the relative difficulty of learning various classifications, and their responses to generalization tests involving novel combinations of cues. The ingredients of the basic network model are shown in Fig. 9.2A.

Presentation of a stimulus or pattern of cues corresponds to activating one or more of the sensory elements on the left. They, in turn, send their activations to an output unit along associative lines that have amplifier weights, the w_i. The weighted inputs are summed at the output node, and this output $\Sigma_{j=1}^{n} w_j a_j$, is converted into a response measure. In classical conditioning, the inputs are single to-be-conditioned stimuli such as lights and bells that are paired with the unconditional stimulus, such as food for a hungry dog; the output node reflects the animal's expectation of the unconditional stimulus given the cues presented. In a classification experiment involving human adults as subjects, the stimuli might be patterns of, say, medical symptoms displayed by a patient, and the output reflects the degree to which the model expects such a patient to have some target disease (classification) versus alternative diseases (Fig. 9.2B).

The network operates in a training environment in which reinforcing feedback (the correct classification) is given after each stimulus pattern. The central axiom of the model is its learning rule, which is that the weights, the w_i's, change on each trial according to

$$\Delta w_i = \beta a_i \left(\lambda - \sum_{j=1}^{n} w_j a_j \right) \tag{4}$$

Here, λ is the training signal which might be +1 if the category is reinforced (e.g., US present) and 0 otherwise (e.g., no US). The cue-intensity parameter, a_i, is assumed to be 1 if cue i is present on the trial, and 0 if it is absent. The learning

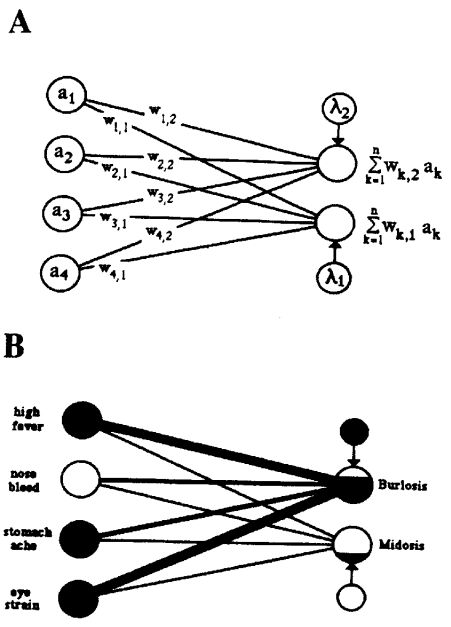

FIG. 9.2. A simple one-layer network that can learn the associations between four stimulus cues and two possible outcomes. (A). The network's classification prediction is a function of the activation on the output nodes. Associative weights between feature nodes and category nodes are updated according to the error-correcting principle of the Rescorla–Wagner (1972) model of classical conditioning, equivalent in this application to Widrow and Hoff's (1960) LMS rule of adaptive network theory. (B). The network applied to a classification experiment involving human adults as subjects, where the stimuli are patterns medical symptoms displayed by a patient and the output reflects the degree to which the model expects such a patient to have some target disease (classification) versus alternative diseases.

rate, β, is a parameter (on the order of .01 in most simulations) that determines how much the weights change when the output differs from the training signal, λ. Equation 4 is variously called the delta rule, the least-mean-square (LMS) rule, or the Rescorla–Wagner conditioning rule (cf. Sutton & Barto, 1981).

Comparing Network and SST Learning Rules

It is instructive to compare Equation 4 of the LMS rule to Equation 3, the linear operator rule from Stimulus Sampling Theory. If we identify p in Equation 3, the probability of responding R_1 with w_1, θ with β, we can re-express the linear operator rule in the terminology of adaptive networks as

$$\Delta w_1 = \beta(\lambda - w_i). \tag{5}$$

Comparing the linear operator rule (Equation 5) of Stimulus Sampling theory with Equation 4 of the Rescorla–Wagner/LMS rule, we note one key difference. Weight changes in the Rescorla–Wagner/LMS rule are governed by the difference (or discrepancy) between the reinforcement (λ) and the network's expectation of the reinforcement, $\Sigma_{j=1}^{n} w_j a_j$ (the output), which is sensitive to all the cues present on a trial. In contrast, Stimulus Sampling Theory operates on each

cue independently; weight changes depend only on the difference between the reinforcement and the current association between cue i and the reinforcing outcome. Note that in training situations where individual component cues are present as complete patterns (as in probability learning studies), Equation 4 of the LMS rule reduces to Equation 3 of Stimulus Sampling Theory. Thus, it is only in training procedures involving patterns of multiple cues that have the opportunity to "compete" among themselves to reduce the error, $(\lambda - \sum_{j=1}^{n} w_j a_j)$, will we expect to see divergent predictions from SST and the LMS network.

Stimulus Sampling Theory follows the tradition of Hull (1943) and Spence (1936) in assuming that the temporal contiguity, or joint occurrence, of a cue and a reinforcing outcome is sufficient for associative learning. This view, however, came under serious attack in the late 1960s, just as interest in Stimulus Sampling Theory began to wane. The work of Kamin (1969), Rescorla (1968), and Wagner (1969) demonstrated that the ability of a previously neutral conditioned stimulus (CS) to become conditioned to an unconditional stimulus (US) depends on the CS imparting reliable, nonredundant, and predictive information about the expected reinforcement. For example, in Kamin's (1969) "blocking" experiment, a light, the CS, was first conditioned to predict a shock, the US. In a subsequent training phase, a compound stimulus consisting of a light and a tone was paired with the shock. Surprisingly, learning of the *tone* → *shock* association hardly occurred at all compared to control subjects who had received no pretraining to the light. One interpretation of blocking and related effects is that animals are learning to modulate the processing of sensory cues in order to adapt out (ignore) the irrelevant cues such as the tone in the given example (Mackintosh, 1975; Pearce & Hall, 1980). These explanations are reminiscent of extensions to Stimulus Sampling that sought to reconcile stimulus generalization with discrimination learning (e.g., LaBerge, 1962; Restle, 1957). Kamin (1969) suggested an alternate interpretation of these attention-like effects. He proposed that the blocking effect results not from modulation of CS-processing but rather from modulation of US-processing. If the effectiveness of a US for producing associative learning depends on the relationship between the CS and the *expected outcome*, little additional learning would occur once the animal had already learned to anticipate (predict) the US (Kamin, 1969).

Rescorla and Wagner provided a precise formulation of Kamin's proposal (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972) and it is this rule that we employ to train the weights in our adaptive network model of human learning. Rescorla and Wagner's conditioning model assumes that the association between a stimulus and its outcome changes on a trial proportional to the degree to which the outcome is unexpected (or unpredicted) given *all* stimulus elements present on that trial (Equation 4). The Rescorla–Wagner model accounts for the blocking effect as follows: When in Phase 1, CS_1 has been initially conditioned to the US, w_1 approaches 1 (assuming $\lambda = 1$ for US trials). If the initial associative strength

of the novel stimulus, w_2, is zero, then the compound stimulus strength, $w_1 + w_2$, will already equal 1 at the beginning of Phase 2. By Equation 4, the incremental change in the associative weight of both stimuli is predicted to be zero when the compound is paired with the US during Phase 2. In contrast to cue-adaptation theories that assume that "attentional" phenomena are mediated by variations in CS processing. Rescorla and Wagner showed how many of these same phenomena could be more readily understood as resulting from variations in US processing.

LMS and the Overlap Problem

Turning back to the "overlap" problem of Stimulus Sampling Theory, we see that the Rescorla–Wagner/LMS rule provides a mechanism for effectively adapting out common irrelevant cues. Consider two stimulus patterns, P 1 and P 2, that are represented by distinct populations of stimulus elements, S_1 and S_2, as well as a common population, S_c. If S_1 is associated with a reinforcing event, R_1, associative strength will accrue to both S_1 and S_c. This association will generalize to P 2 via the overlapping elements in S_c that are shared with P 1.

In a discrimination training procedure, however, P 1 might be associated with R_1 and P 2 with R_2. One possible network representation of this problem is to have a single output node that receives a training signal of $+1$ when R_1 is reinforced and a training signal of -1 when R_2 is reinforced. Under these conditions the competitive learning principle of the Rescorla–Wagner/LMS rule will seek a solution whereby $w_1 + w_c = +1$ while $w_2 + w_c = -1$. One possible solution is to have all of the associative strength accrue to $w_1 = +1$ and $w_2 = -1$, with $w_c = 0$ "adapting out" so that the system achieves errorless discrimination (see also Rudy & Wagner, 1975, p. 290). As we see later, however, it is possible under some training procedures for the LMS network to find other solutions that do not require that $w_c = 0$. A major challenge for the LMS network—and all models of learning—is to try and reconcile the role of common elements in both stimulus generalization and discrimination learning.

LMS and Probability Matching

Like Stimulus Sampling Theory, the LMS network will generally predict probability matching in choice behavior when the output activations (or a mono tonic transformation of them) are converted to choice probabilities using a likelihood ratio rule (Gluck & Bower, 1988a). The relationship between the Least Mean Squares solution and probability matching can most easily be seen with reference to a single output node that is reinforced ($\lambda = 1$) with probability π. If A is the output activation of the node, then the squared error will be $(1 - A)^2$ with probability π and A^2 with probability $(1 - \pi)$. Thus, the expected mean squared error (MSE) is

$$E[MSE] = \pi(1 - A)^2 + (1 - \pi)A^2. \quad (6)$$

To find the value of A that minimizes the expected mean squared error, we differentiate Equation 6 with respect to A:

$$\frac{d(E[MSE])}{dA} = -2\pi(1 - \pi) + 2(1 - \pi)A. \quad (7)$$

By setting $d(E[MSE]) = 0$ we solve for A to find that the minimum squared error occurs when $A = \pi$ (see also Gluck & Bower, 1990, p. 108). Thus, we expect that the LMS algorithm will converge to a set of weights that result in the closest possible approximation to having the output activations reflect the observed probabilities of reinforcement for each pattern in the training set. Because the approximation to probability matching on the output activations is not necessarily bounded between 0 and 1 (as are the p in Stimulus Sampling Theory) it may be necessary to transform the network's output activations before mapping them onto expected choice probabilities. Examples of using the LMS network to fit observed data on probability matching can be found in Gluck and Bower (1988a), Estes, Campbell, Hatsopoulos, and Hurwitz (1989), and Shanks (1989).

Binder and Estes (1966): An Adaptive-Network Interpretation

We return now to the Binder–Estes study to see what the LMS network predicts here. Medin and Edelson (1988), in their replication and extension of the Binder–Estes study, noted that the "relative-novelty" effect is qualitatively consistent with competitive learning rules, such as the Rescorla–Wagner rule. Their logic goes as follows: Assume that cues a and b compete to predict the common category while cues a and c compete to predict the rare category. Because a occurs more often with the common rather than the rare category, it will presumably acquire more associative weight to the common category. Thus, a will compete with b to predict the common category, thereby diminishing b's association to the common category. For pattern ac to predict the rare category, symptom c will have to overcome the association of a to the common category. This leads us to expect that when b and c are paired together, c's association to the rare category should be stronger than b's association to the common category. Thus, a competitive-learning principle might expect that the novel test pattern bc should be judged more strongly associated with the rare category, as observed by Binder and Estes (1966) and Medin and Edelson (1988).

Given this reasoning, we might expect that the LMS network model in Fig. 9.2, which incorporates Rescorla and Wagner's competitive learning rule, should account for the relative-novelty effect. However, as Medin and Edelson (1988, p. 75) note the Rescorla–Wagner model predicts that with extended training, b and c will accrue all the associative strength, leaving a with none. Figure 9.3 shows an adaptive network model of the Binder–Estes/Medin–Edelson experiment. The network has three input nodes: one for each of the three symptoms. All

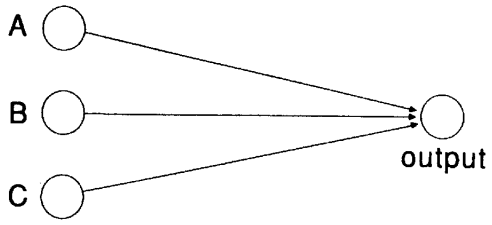

FIG. 9.3. An adaptive network model of the Binder–Estes task using a local representation scheme with one input node for each of the three stimulus components and a single output node that receives reinforcement of +1 for the more common outcome and −1 for the less common (rare) outcome. This one-output node model is equivalent in this application to a two-output node model in which each output nodes receives reinforcement of 1 (category present) or 0 (category not present).

weights are initialized at 0. The presence or absence of cue-i is represented by an input node activation, a_i, of 1 or 0, respectively. The output node is reinforced with $\lambda = +1$ for the common category and $\lambda = -1$ for the rare category. This one-output-node model, with $+1/-1$ reinforcements, yields identical predictions to a two-output-node model with 1/0 reinforcements where each output node corresponds to one of the possible outcomes (see Gluck & Bower, 1988a, footnote 2, p. 234, for more details on this correspondence).

Figure 9.4 graphs the changes in weights for the three input nodes (cues) during training, the output activations for the training patterns during learning, and the output activations (responses) for the transfer patterns at each stage in learning. These simulations are from a network run for 200 trials with a learning rate, β, of .03; so long as β is sufficiently small, however, the important ordinal predictions of the model are independent of the particular parameter value chosen. The simulation in Fig. 9.4 confirms Medin and Edelson's observation that extended training with the Rescorla–Wagner/LMS rule results in cues b and c acquiring all the predictive strength: asymptotically, $w_B = +1$, $wc = -1$, whereas cue a adapts out, with $w_A = 0$. Thus, b is completely associated with R_1, the common category, c is completely associated with R_2, the rare category, and a has no associative strength at all. As shown in Fig. 9.4A, a does acquire a pre-asymptotic association to the common category (i.e., a positive weight). Thus, the network's early response to pattern a is consistent with both Binder and Feldman's (1960) and Medin and Edelson's (1988) results on the common-cue test (a). Medin and Edelson (p. 75) and subsequently we (Gluck & Bower, 1988a) incorrectly suggested that the relative-novelty effect will also emerge from the Rescorla–Wagner model as a pre-asymptotic effect. As the simulations in Fig. 9.4 demonstrate, this is clearly incorrect. Figure 9.4C shows that the network's response to the transfer pattern, bc, favors the common category at all stages of learning prior to asymptotic learning. Because of the imbalance in presentation frequencies, the response to bc remains positive despite a's transient association to the common category, because w_b increases in strength towards $+1$ much faster than the w_c approaches -1. Early in training, during the tran-

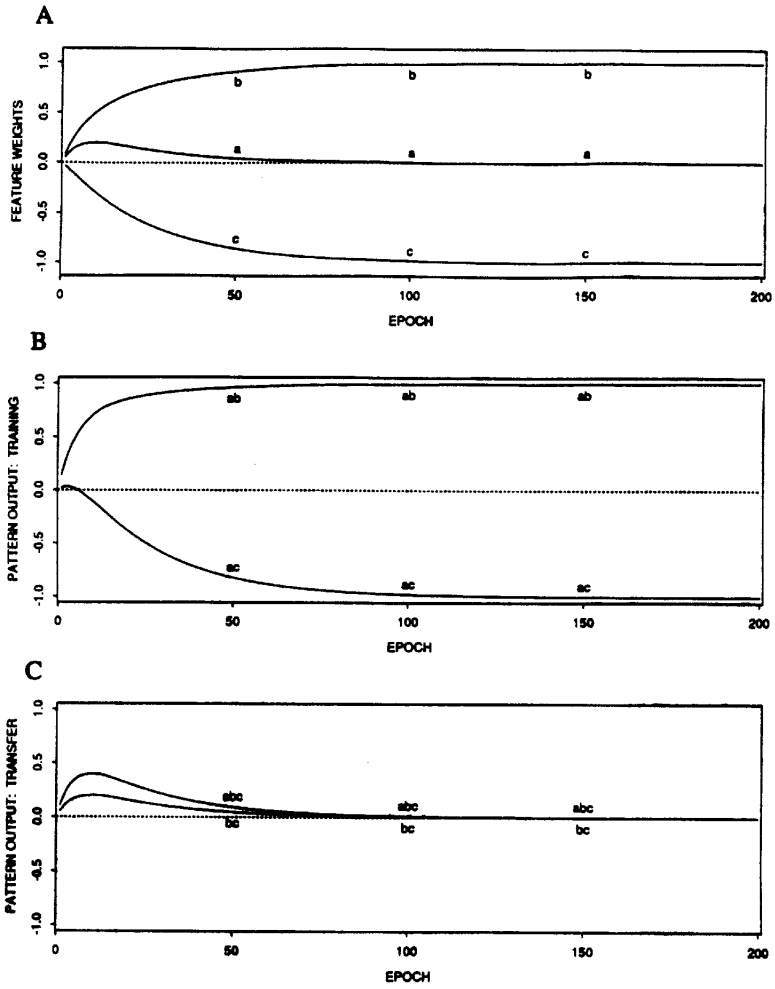

FIG. 9.4. Gluck and Bower's adaptive network model applied to Binder–Estes/Medin–Edelson experiments. (A). Changes in weights for the three input nodes (cues) during training across training. Positive weights and activations favor the common category whereas negative weights and activations favor the rare category. (B). Output activations for the training patterns during training. (C). Predicted responses to transfer tests at each point in training. In all these figures, positive weights and activations favor the common category whereas negative weights and activations favor the rare category.

sient association of *a* to the common category, *b*'s association to the common category is always greater than *c*'s association to the rare category. Thus, at no time during the course of training will the Rescorla–Wagner model—or, equivalently, the adaptive network model of Gluck and Bower (1988a)—predict the "relative novelty" effect on the *bc* test.

Understanding the Network's Solution

Why does the network's behavior in this training situation differ from our intuitive expectation of what a competitive learning rule should do? When the network is trained as just described, the LMS rule converges on the "solution vector," $W_{[a,b,c]} = [0,+1,-1]$. Note that this is only one of many possible solution vectors that would be equally effective in solving the *ab*/*ac* discrimination. For example, if $W_{[a,b,c]} = [.2,.8,-1.2]$, this would also result in errorless performance. In a deterministic task that can be perfectly solved by the network (i.e., MSE = 0), the set of solution vectors is unaffected by variations in the presentation frequencies of the individual training patterns. This type of problem, for which multiple solutions exist, can be contrasted with other discrimination problems that have unique solutions. For example, the nondeterministic classification task in Experiment 1 of Gluck & Bower (1988a, Appendix A) has a unique solution that can be derived analytically. When a unique network solution exists, the LMS algorithm will converge on that solution independent of the initial weights, assuming a sufficiently small learning rate (Widrow & Hoff, 1960). In situations where multiple solutions exist, such as the Binder–Estes/Medin–Edelson task, the final weights obtained with the Rescorla–Wagner/LMS algorithm will be sensitive to their initial values (Gluck & Bower, 1988b, Appendix B; Parker, 1986). The sensitivity of the LMS rule to initial conditions is familiar to animal learning theorists as the property that allows the Rescorla–Wagner model to account for the effect of pretraining in Kamin's (1969) blocking study.

If many different solutions are equally "good" in minimizing the expected squared error, why does the network converge to $[0,+1,-1]$ rather than another solution, for example, $[.2,.8,-1.2]$? To see why, it is helpful to consider the set of all possible solution vectors as being a subset of the three-dimensional "weight space" that characterizes all possible states of the three-weight network. If the network begins with all weights set to zero, then the solution with the smallest sum squared weights represents the "closest" solution to the initial conditions, where closeness is measured by Cartesian distance. Parker (1986) has shown that if the weights in the network are initialized at zero (or randomly distributed with zero mean), the asymptotic weights will tend toward the solution closest to the initial conditions. For the network model of the Binder–Estes/Medin–Edelson task we expect, on average, a solution where cues "b" and "c" have all the weight because the solution to the simultaneous linear