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

$$p_{n+1} = \begin{cases} (1 - \theta)p_n + \theta & \text{if } \mathrm{R}_1 \text{ is reinforced} \\ (1 - \theta)p_n & \text{if } \mathrm{R}_2 \text{ is reinforced.} \end{cases} \qquad (1)$$

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

$$
\begin{aligned}
p_{n+1} &= \pi[(1-\theta)p_n + \theta] + (1-\pi)(1-\theta)p_n \\
&= (1-\theta)p_n + \theta\pi \qquad (2)
\end{aligned}
$$

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 $\left(\frac{1}{T}\sum_{n=1}^{T} \mathrm{p_n}\right)$, 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 $n + 1$,

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

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).

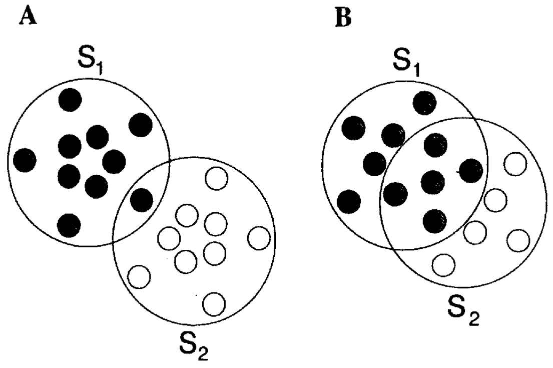

FIG. 9.1. Generalization through shared elements for two stimuli S_1 and S_2 that are: (A) slightly similar (small overlap) or (B) very similar (large overlap).

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 information 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

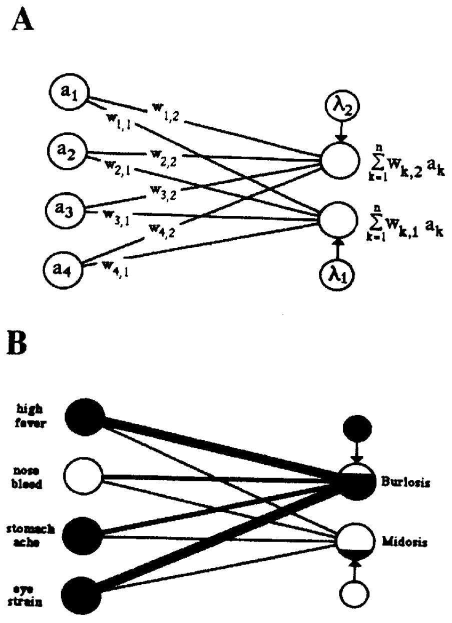

FIG. 9.2. A simple one-layer network that can learn the associations between four stimulus cues and two possible outcomes. (A). The network's classification prediction is a function of the activation on the output nodes. Associative weights between feature nodes and category nodes are updated according 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

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

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

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

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

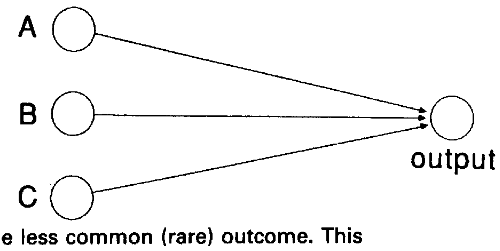

FIG. 9.3. An adaptive network model of the Binder–Estes task using a local representation scheme with one input node for each of the three stimulus components and a single output node that receives reinforcement of +1 for the more common outcome and −1 for the less common (rare) outcome. This one-output 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-

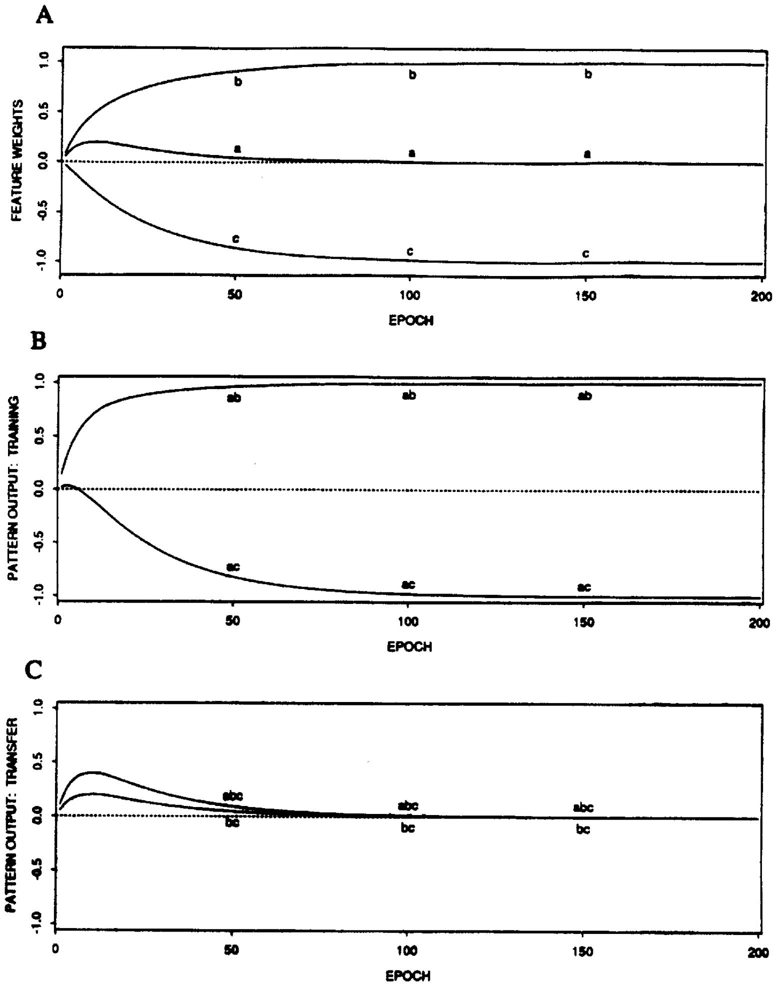

FIG. 9.4. Gluck and Bower's adaptive network model applied to Binder–Estes/Medin–Edelson experiments. (A). Changes in weights for the three input nodes (cues) during training across training. Positive weights and activations favor the common category whereas negative weights and activations favor the 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 equations $w_A + w_B = +1$ and $w_A + wc = -1$, with the smallest sum-squared weights is $w_A = 0$, $w_B = +1$, $wc = -1$. (See Gluck & Bower, 1988b, Appendix B, for more details on deriving the expected asymptotic convergence when multiple solutions exist).

STIMULUS SAMPLING AND THE RESCORLA-WAGNER MODEL

It is clear from the analyses presented that neither Stimulus Sampling Theory nor our LMS network model provides an adequate account of the effects of category frequency on discrimination learning and transfer generalization when the training patterns share common cues. Stimulus Sampling Theory accounts best for transfer effects involving single component cues. This "component matching" principle was summarized by Binder and Tayler (1969) as: "If two or more different responses have been reinforced in the presence of a given cue during training, then with any later tests the probability that any one of these responses will be evoked by the given cue is equal to its relative frequency of reinforcement" (p. 91). In contrast, adaptive network theory, and the Rescorla–Wagner/LMS rule in particular, provides a better account of discrimination learning when the reinforcement contingencies for cues are dependent on the context in which they appear. The LMS rule converges on a set of weights that (as closely as possible) produce "pattern matching" probabilities as activations on the output nodes when the individual cue weights are combined additively. To paraphrase Binder and Taylor: The LMS network seeks a solution whereby if two or more different responses have been reinforced in the presence of a given *pattern* during training, then with any later tests the output activation evoked by the given *pattern* for one of these responses will be equal to its relative frequency of reinforcement.

A possible rapprochement between the explanatory abilities of Stimulus Sampling Theory and the Rescorla–Wagner model has been suggested by Rescorla (1976) and Blough (1975). These authors have shown that integrating the learning rule from Rescorla–Wagner's (1972) conditioning model with the stimulus-representation assumptions from Stimulus Sampling Theory can account for several animal learning behaviors. We first review the applications of this hybrid model to animal learning behavior and then consider its implications for human classification learning.

Rescorla (1976)

Rescorla (1976) highlighted an important implication of the Rescorla–Wagner model when it is applied to a distributed representation in which similar stimuli share common features. Traditional learning theory says that the optimal way to train an associative connection is by direct reinforcement of the target stimulus. Rescorla, however, reported a paradoxical case where training to a *generalized* stimulus enhances performance to a target stimulus more than reinforcement to the target stimulus itself. This paradoxical outcome and the circumstances that produce it are predicted by Equation 4 of the Rescorla–Wagner/LMS rule.

If we conceptualize the similarity of two stimuli in terms of common or shared stimulus elements, then we may represent even a simple stimulus like a high-pitched pure tone as a compound of stimulus elements, denoted AX; another similar stimulus like a low-pitched pure tone would be represented as another compound, BX. Here, X denotes the set of common elements, whereas A and B denote those sets of sensory elements unique to the two stimulus sets. In conditioning the high tone (AX) to a shock US, Equation 4 applies to the separate A and X components of that stimulus. With repeated reinforcement of AX, the weights w_A and w_X will increase together until their sum equals the reinforcing value of $\lambda = 1$ where each might be, say, about ½. Were we to continue to reinforce AX beyond this point, the LMS rule of Equation 4 expects no change in the strengths of the A and X associations. Now consider what would happen if we gave trials wherein a generalized stimulus (the low tone, denoted BX) was paired with shock. Because B begins at low strength, the combination BX begins with an association strength far below 1. During a block of reinforced trials on BX, Equation 4 implies that w_B and w_X will increase. This increase in w_X should be most apparent when we test the subject again on the original training stimulus, AX. On such a test, the compound strength $w_{AX} = w_A + w_X$ will be higher than before, higher even than if the subject had just continued training on AX alone. In an experiment of this kind, Rescorla (1976) found just this result. Training to a generalized stimulus (following initial learning) produced greater conditioned responding to a target stimulus than did extended training on the target stimulus itself. This is a most counterintuitive result, and one that provides impressive support for the Rescorla–Wagner learning rule when combined with a common-elements representation of stimulus similarity.

Blough (1975)

Nearly concurrently with Rescorla (1976), Blough (1975) described a stimulus sampling model that incorporated a generalization of the Rescorla–Wagner rule very similar to Widrow and Hoff's (1960) LMS rule. Blough assumed that a stimulus continuum (such the pitch of a tone or the wavelength of light) could be represented as an ordered sequence of overlapping sets of hypothetical stimulus elements. Presentation of a physically-defined stimulus corresponds to sampling a subset of these elements according to a unimodal bell-shaped probability distributed with mode at the internal unit corresponding to the physical stimulus. Thus, different physical stimuli are assumed to project probabilistically to internal sensory units that overlap to varying degrees, as illustrated in Fig. 9.5.

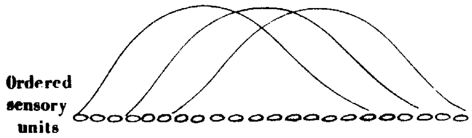

FIG. 9.5. Probability distributions for sampling sensory units given the presentation of physical stimuli at three levels along a stimulus continuum.

Consider now what happens if stimuli below some point on the continuum are reinforced at one rate (e.g., on one twelfth of the trials) while stimuli above that point are reinforced at another rate (e.g., on one third of the trials). What does the Rescorla–Wagner/LMS rule predict will happen with this differential training? The predicted asymptotic association strengths for each stimulus element are shown as a dashed line in Fig. 9.6A. As expected, the model predicts that subjects should adjust their conditional associations to reflect the two reinforcement levels. The distributed representation of the physical stimuli results in a smooth transition from one level of responding to the other because elements nearby the transition point are activated by physical stimuli both above and below the transition point. This smooth transition is also predicted by the linear operator rule from SST when applied to the same stimulus representation.

More interesting than the gradual transitions from one level of responding to the other are the exaggerated "shoulder" and "trough" on either side of the transition point in Fig. 9.6A. These predicted contrast effects are analogous to edge-enhancement (so-called "Mach bands") in sensory psychophysics. The unique prediction of the Rescorla–Wagner model is that the elements just a little bit further away from the transition point should become "superconditioned" because they frequently co-occur—and hence compete for associative strength—with near-edge elements whose associative strengths reflect conditioning at mixed levels of reinforcement.

The plausibility of such contrast effects in discrimination learning were demonstrated by Blough (1975) who trained hungry pigeons to peck a colored key for food. Keypecks at wavelengths below 597nm were reinforced on one twelfth of the trials, whereas keypecks at wavelengths above 597nm were reinforced on one third of the trials. As predicted by Blough's Stimulus-Sampling extension of the Rescorla–Wagner model, the animals' pecking rates showed marked shoulder and trough effects whereby the cues that were just above and below the transition point for reinforcement (597 nanometers) appear to be "superconditioned" beyond the steady-state levels associated with wavelengths further away from 597 nanometers (Fig. 9.6).

A DISTRIBUTED 'STIMULUS SAMPLING' NETWORK MODEL

An adaptive network model in which stimuli are represented by stochastically activating overlapping populations of input nodes (stimulus elements) is one type

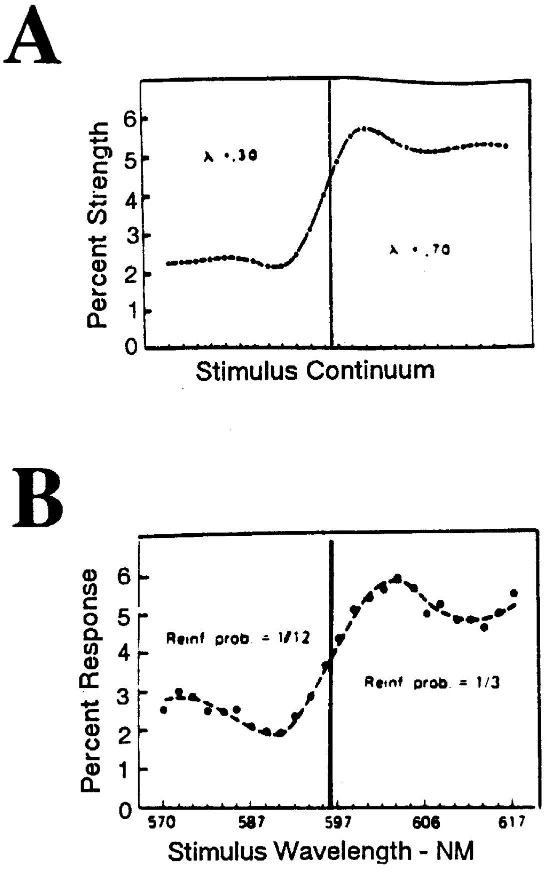

FIG. 9.6. (A). Predicted asymptotic strengths from the Rescorla–Wagner rule for discrimination training on a stimulus continuum. (B). Response rates of three pigeons to key lights varying in wavelength. The vertical line separates the high and low reinforcement stimuli. Note the trough and peak to the left and right of the edge. (From Blough, 1975).

of "parallel distributed network" (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). Networks like this that embody a *distributed representation* use entire patterns of activation across many units to represent different concepts, with different patterns of activation corresponding to different concepts or features (Hinton, McClelland, & Rumelhart, 1986). This is in contrast to a *local representation,* in which each unit (node) in the network is taken to represent a single concept or feature (Feldman, 1985). This "local" representation is what we have previously adopted in our models of human learning, as illustrated in Fig. 9.2B.

A key feature of distributed representations is that each unit is involved in representing many different concepts. Whereas local representation schemes are conceptually easier to understand, many network theorists have been led to adopt distributed representations because of some of their interesting emergent properties. The most compelling advantage of a distributed representation is its ability to generalize automatically from previous training to similar novel situations. As in the earlier Stimulus Sampling Theory, generalization emerges in distributed representations because similar conceptual entities are encoded by activating overlapping sets of units. Several researchers have used this property of distributed networks to account for various aspects of cognitive functioning (Anderson, Silverstein, Ritz, & Jones, 1977; Hinton & Anderson, 1981). Another appeal of distributed representations is the compelling intuition that they are more biologically plausible than local representation schemes (Lashley, 1929; Sejnowski, 1988; Thompson, 1965), but we focus here solely on the behavioral implications of distributed representations.

Although many of the generalization properties of distributed representations bear a marked resemblance to human behavior, there has been little attempt to apply these principles to fitting precise details of human learning and generalization. Part of the problem has been that there is little consensus among theorists as to how external stimuli should be identified with distributed patterns of activity. Building on the previous successes of Stimulus-Sampling-Theory's stimulus representation provides a possible formalism for developing a "distributed" network theory of psychology representation.

Binder and Estes (1966): A Distributed-Network Interpretation

Incorporation of Stimulus Sampling into the network model requires consideration of two new factors. First, it adds an element of randomness or stochasticity to our representation of the stimulus conditions operating on a trial. In the previous "local" network model (Gluck & Bower, 1988a, 1988b), the functional representation of stimuli was identical to the nominal stimuli as described by the experimental paradigm. In stimulus-sampling, the functional representation is presumed to include only a random subset of the nominal stimulus conditions. The second factor introduced by a sampling representation is the explicit incorporation of stimulus similarity through the activation (sampling) of common stimulus elements.

To better understand the unique implications of these two factors, we begin by considering a "non-overlapping" model, in which the population pools for the three cues, *a*, *b*, and *c* are distinct and have no common elements. The behavior of this model will address the implications of adding stochasticity to our model, independent of the effects of common-elements. After analyzing the behavior of this non-overlapping model, we consider an "overlapping"model, which incorporates stimulus similarity among the three cues through the activation of common elements.

Stochasticity in Stimulus Representation

We begin by assuming three distinct pools of input nodes, each having n elements (Fig. 9.7). Presentation of a stimulus cue is presumed to stochastically activate the elements in the corresponding pools with probability θ. Thus, on average, we can expect $n\theta$ elements to be activated in each population of elements. Figure 9.8A shows the result of this non-overlapping network model applied to the Binder–Estes/Medin–Edelson experiments. In Fig. 9.8C we see that the response to the transfer pattern *bc* has an initial upward swing due to the unequal presentation frequencies of the two categories. At asymptote, however, the associative strength of *c* for the rare category outweighs the strength of *b* for the common category and the model correctly predicts that the compound conflicting test, *bc*, will favor the rare category. Thus, the addition of a *stochastic sampling processes* to the network provides a formal instantiation of Medin and Edelson's qualitative proposal for how a competitive learning rule can account for the "relative novelty" effect.

Why does the relative novelty effect emerge in these simulations and not in a "local representation" network? Note that the formal properties of the network have not changed. What has changed is only our assumption about how the external world is represented on the input nodes. If the sampling rate θ is set to 1, the distributed model reduces to the local model. The important consequence of

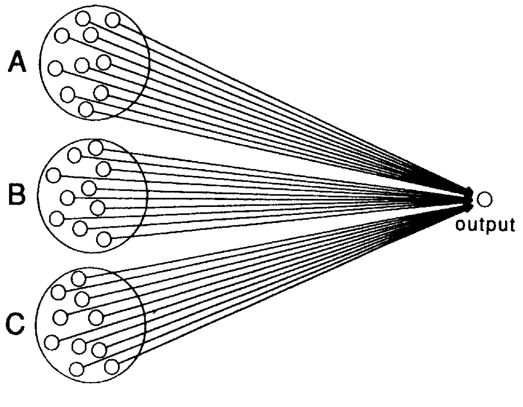

FIG. 9.7. An adaptive network model of the Binder–Estes task using a representation scheme with one pool of input nodes for each of the three stimulus components.

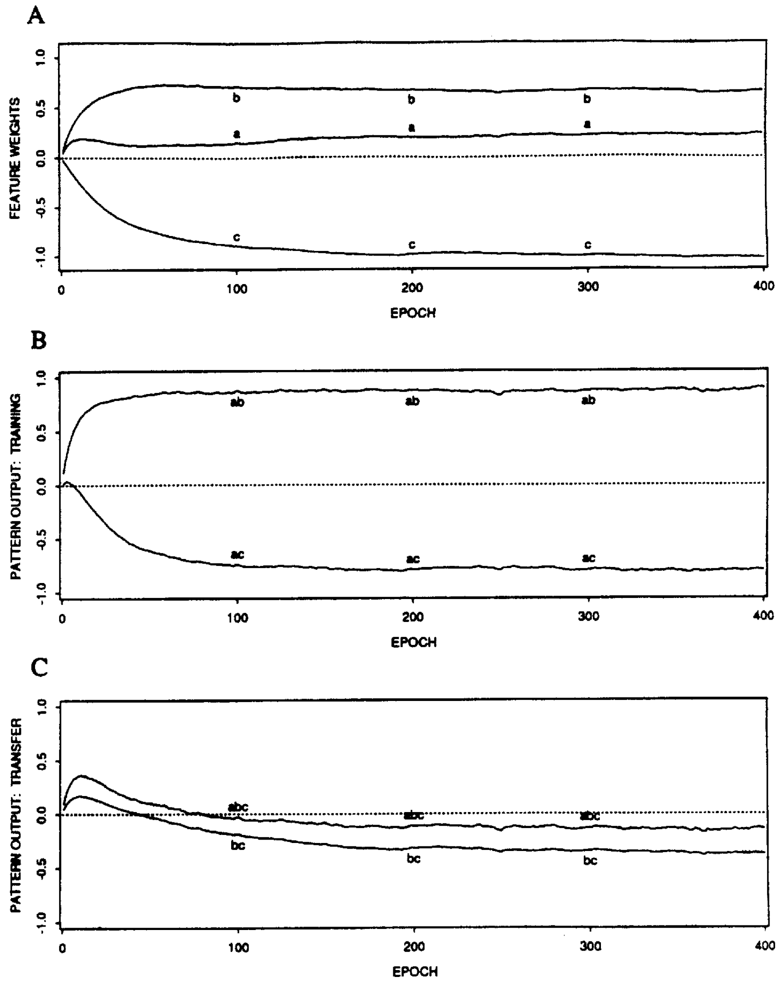

FIG. 9.8. A distributed network model applied to the Binder–Estes/Medin–Edelson experiments with $\beta = .03$, $n = 20$ and $\theta = .2$. (A) shows the expected summed weights (for a given trial) for weights from each of the three input nodes (cues) during training. Positive weights and activations favor the common category whereas negative weights and activations favor the rare category. (B) graphs the expected output activations for the training patterns during training. (C) shows the expected responses to transfer tests at each point in training. Parameter values used were: $\beta = .03$, $n = 20$, and $\theta = .2$. As in Fig. 9.4, positive weights and activations favor the common category whereas negative weights and activations favor the rare category. Each graph is the result of averaging many simulations of these conditions.

the change in representation is that it alters the nature of the discrimination from a deterministic problem to a probabilistic problem. With a stochastic representation of the stimulus environment, there exists a unique set of weights that provide a least mean squares solution. In situations where there is a unique solution, the distributed and local network models will make effectively equivalent predictions. For example, we have analyzed the stimulus environments presented in Experiments 1 and 2 from Gluck and Bower (1988a), and the important ordinal predictions of the original model are maintained by the sampling network. A more detailed exposition of the translation from local to distributed representations, and the conditions under which they make equivalent predictions, can be found in Stone (1986).

Because the processing assumptions of this SST-Network model are identical to those of the previously described network, we can use the same analytic tools for deriving asymptotic solutions (see Appendix A, Gluck & Bower, 1988a). To do so requires making assumptions about n and θ (β has no effect on the expected asymptotic weights). With the values of $n = 20$ and $\theta = .2$ used in the simulation in Fig. 9.8, the expected asymptotic weights for the individual nodes in the three pools are $w_A \approx .057$, $w_B \approx .160$, $wc \approx -.256$. An average of $n\,\theta = 4$ nodes will be active on each trial in each pool, so the expected activations resulting from the presence of each of the component cues alone is .227, .644, and -1.02, for a, b, and c, respectively.

An intuitive explanation of the relative novelty effect is that the sampling representation adds an additional constraint to the search for an appropriate solution. The constraint can be loosely characterized as *robustness* and is a direct consequence of introducing "noise" into the training procedure. The network now searches for a solution that not only solves the $ab \rightarrow R_1/ac \rightarrow R_2$ discrimination, but is also maximally tolerant of noisy information about cues. For example, if we knew that the first feature was a but were unsure about the second feature, we would want the network to prefer the common category. The Rescorla–Wagner/LMS rule, by virtue of its competitive nature, searches for a parsimonious solution in which redundant information is ignored. But a parsimonious solution may be brittle in that it requires complete and perfect information about all features in the stimulus pattern. When multiple solutions are equally valid, the addition of stochastic noise biases the system toward solutions that are noise tolerant. Similar results have been found for learning in multilayer adaptive networks (Elman & Zipser, 1988; Hanson, 1990). We noted earlier that the LMS rule converges on weights that approximate probability matching on stimulus *patterns* in contrast to the linear operator rule from SST that converges on weights that result in probability matching on *component cues*. The addition of stochasticity to the network provides an additional constraint: When multiple solutions exist that are equally effective in producing "pattern matching," the network will prefer the solution that best approximates "component matching."

The addition of the stochastic sampling representation changes a deterministic discrimination into a probabilistic discrimination in which the expected squared error can never be totally reduced to zero. This does not imply, however, that choice performance, as measured by expected percent correct, cannot reach 100%. In contrast to Stimulus Sampling Theory, which directly maps associative weights onto response probabilities, the network model is not committed to a specific response mapping rule. Rather, it assumes that an unspecified monotonic rule converts activations into response probabilities. For example, in Gluck and Bower (1988a,b), we adopted the sigmoidal transform, which has one free parameter describing the gain or slope of the "S-shaped" sigmoid. As long as all input patterns are more associated with their correct response than with any incorrect response, the sigmoidal response rule can bring choice performance arbitrarily close to 100%, depending on the gain constant. Even with a more moderate gain constant, the sigmoidal transform has the effect of compressing activations near the boundary values so that output activations from all large weights are near unity.

Limitations of the Non-overlapping Representation

One problem with the predictions in Fig. 9.7 is that the model fails to account for Medin and Edelson's finding that subjects, under some conditions, judge the pattern *abc* to be more strongly associated with the common category. We now consider the effects of incorporating stimulus overlap, as well as stochasticity, in our representation of the stimulus cues and show how this effects the predicted response to the *abc* pattern.

With three cues, S_A, S_B, and S_C, we could model the overlapping elements shared by S_A and S_B, S_A and S_C, S_B and S_C, and those shared by all three cues as shown in Fig. 9.9. A simpler alternative, which we adopt here, is to assume that presentation of a cue results not only in the activation of nodes in its "own" pool (with probability θ), but also in the activation of nodes in other pools (with probability θ').

As θ' approaches θ, the weights (and output activations) become more biased towards the presentation frequencies of the categories. Note that in the extreme case in which $\theta = \theta'$, the network has no discriminative input, only random noise on the input nodes, and the input activations would be uncorrelated with the stimulus environment. In this case, the distributed network model reduces, in all aspects, to the Stimulus Sampling Theory account of probability matching described earlier. Clearly $\theta' \ll \theta$ if the cues are reasonably discriminable. Choosing the same values for θ and n that we used in the simulation in Fig. 9.8, and setting $\theta' = .07$, the results of the new simulation are shown in Fig. 9.10. By comparing Fig. 9.10C with Fig. 9.8C, we see that the incorporation of a common-elements representation of stimulus similarity alters the model's behavior so that the correct predictions are obtained for both the relative-novelty test ($bc \rightarrow R_2$) and the combined test ($abc \rightarrow R_1$). The degree of overlap, $\frac{\theta'}{\theta}$, will determine the relative strength of this combined test result.

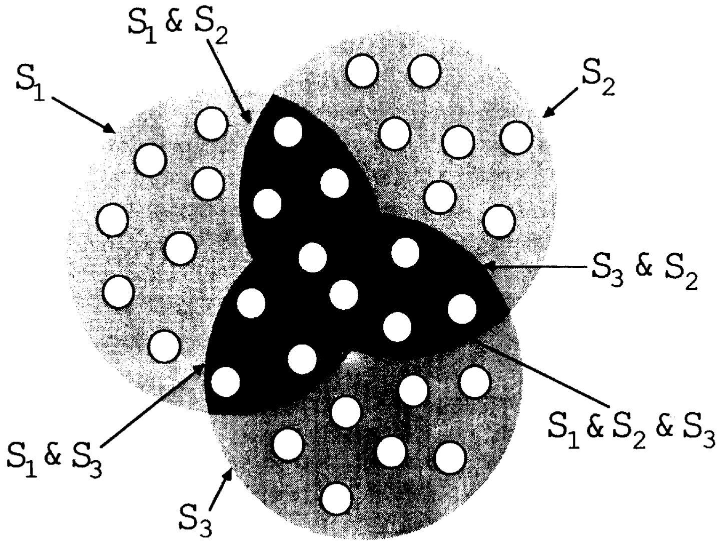

FIG. 9.9. One method for having three overlapping pools of stimulus elements corresponding to three stimulus cues, S_1, S_2, and S_3.

The model also predicts that both the *bc* and *abc* patterns should show a preference for the more common outcome (R_1) very early in training, with the *bc* pattern reversing its preference with further training. No data is currently available, however, to test this "crossover" prediction.

In comparing our account of the classification of the relative-novelty (*bc*) and the combined patterns (*abc*), it is important to note that the explanation of the former is parameter free ($0 < \theta < 1$). Thus, we expect the "relative novelty" effect to be strong and reliable. Indeed, it has been replicated by various investigators over the last 20 years (Binder & Estes, 1966; Binder & Taylor, 1969; Medin & Edelson, 1988; Medin & Robbins, 1971). In comparison, our account of the combined test (*abc*) depends critically on the "stimulus confusability" of the component features as measured by the magnitude of θ' relative to θ. Here we can only make the weaker claim that the model is sufficient to account for the *abc* → common preference. However, this dependence on the relative magnitude of θ' suggests that experimental manipulations designed to influence stimulus confusability might change the results of the combined test. One such manipulation was used by Medin and Edelson in their fourth experiment. They gave subjects one of two types of instructions. In the focus condition, subjects were told to focus on the symptoms that proved most reliable. In the complete condition subjects were told that they should learn about all of the symptoms. If we assume that the focus condition decreased the opportunity for stimulus confusion, effectively lowering θ', then we expect subjects in the focus condition to exhibit a stronger "relative novelty" effect on the conflicting test (*bc*), but less bias for the

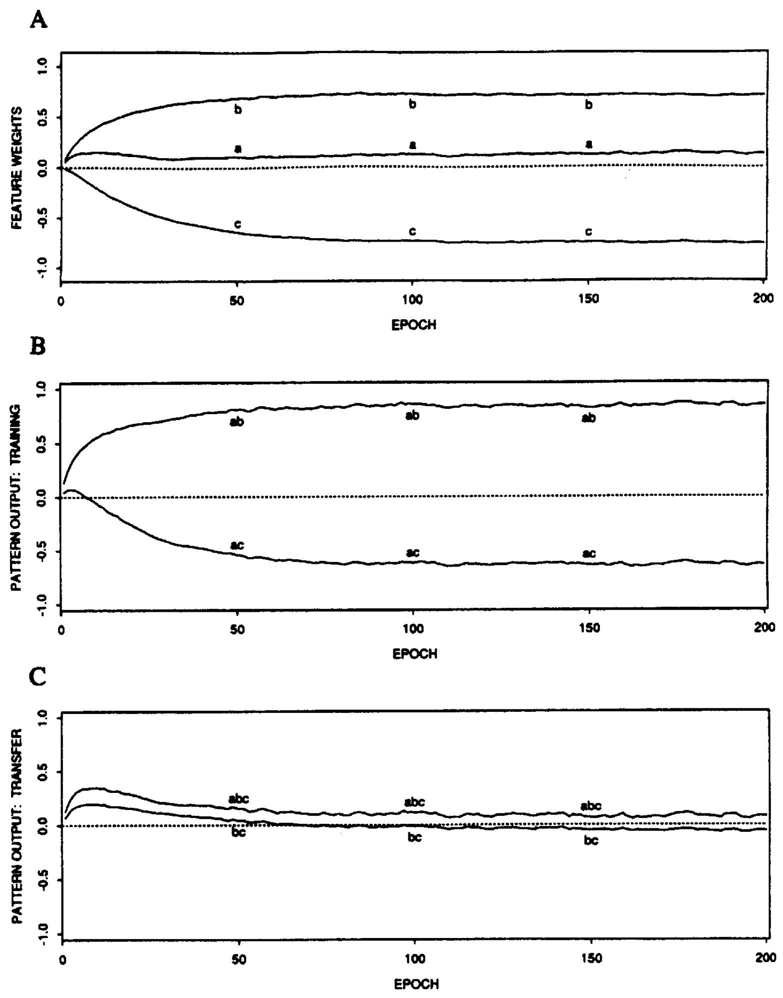

FIG. 9.10. A distributed network model applied to the Binder–Estes/Medin–Edelson experiments with $\beta = .03$, $n = 20$, and sampling probabilities of $\theta = .2$ and $\theta' = .07$. (A) shows the expected summed weights (for a given trial) for weights from each of the three input nodes (cues) during training. Positive weights and activations favor the common category whereas negative weights and activations favor the rare category. (B) graphs the expected output activations for the training patterns during training. (C) shows the expected responses to transfer tests at each point in training.

common category on the combined (*abc*) tests. This is precisely what Medin and Edelson found. On the conflicting tests, the focus group showed a considerably larger bias for the rare category than did the subjects in the complete group. On the combined test, subjects in the complete condition exhibited the same effect found in the other experiments—they showed a bias for the common category. Subjects in the focus group, however, showed the opposite effect, exhibiting a slight preference for the rare category versus the common category. Thus, when given the "focus group" instructions, subjects behaved much like the network in Fig. 9.8 with an effective θ' of 0.

Summary

Binder and Estes (1966) were able to account for some of their data within the framework of Stimulus Sampling Theory. They were unable, however, to account for two phenomena. First, subjects discriminate perfectly patterns that share common features. Second, subjects exhibit a relative novelty effect in which two conflicting cues are more strongly associated with a less frequent outcome. Following Rescorla (1976) and Blough (1975), we have developed a distributed network model that combines the learning rule from Rescorla and Wagner's (1972) conditioning model with the stimulus representation assumptions from Stimulus Sampling Theory. This distributed network predicts Binder and Estes' relative-novelty effect (*bc* → less frequent outcome), component matching on transfer (*a* → more frequent outcome), and provides a possible account for Medin and Edelson's (1988) combined-pattern results (*abc* → either, depending on training instructions).

These analyses suggest that it may not be appropriate to apply the Rescorla–Wagner/LMS rule directly to a veridical representation of stimuli, especially for deterministic discriminations. In discrimination tasks in which an infinite number of least-squares optimal solutions exist, the "local" network will often find a solution that is simple and nonredundant but extremely sensitive to "noise" or perturbation. These solutions do not yield transfer generalizations in accord with empirical data. Training the network with a stochastic representation of the input stimuli results in a "robust" solution that generalizes more effectively. Stimulus Sampling Theory provides a stochastic formalism for representing input stimuli.

Configural Cues and Stimulus Sampling

In these preliminary analyses of a distributed network model, we have represented stimulus patterns as collections of component cues (e.g., *a*, *b*, and *c* in the Binder–Estes/Medin–Edelson experiments). In other work (Gluck, in press; Gluck, Bower, & Hee, 1989) we have extended this component representation to include pair-wise conjunctions of features as unique cues. This "configural-cue" model accounts for several aspects of complex human category learning and animal learning. We have not included these "higher order" cues in our analyses of the Binder and Estes and Medin and Edelson experiments, because the relevant configural-cues, *ab* and *ac*, are redundant in this application with the component-cues, *b* and *c*. The consideration of configural-cues does not change the basic predictions of the network model for the Binder–Estes/Medin–Edelson discrimination task. In general, however, the configural-cue approach is perfectly compatible with a distributed stimulus-sampling representation. Configural-cues, like component-cues, can be represented as populations of elements. Furthermore, the geometric interpretation of stimulus similarity as shared common elements (Figs. 9.1 and 9.9) can be extended to represent configural-cues. When overlapping elements are activated only by the presence of both cues (rather than by either alone), they represent unique configurations of the component cues, that is, configural-cues.

CONCLUSION

In describing the strategy for theory development that guided the growth of Stimulus Sampling Theory, W. K. Estes (1982) wrote:

> The approach I favored [was] starting with associative concepts already established for simpler forms of learning and progressively modifying and elaborating them as successive approximations to an adequate theory are confronted by new facts. . . . My preferred strategy was not to discard the original concept but rather to extend it to a broader conception of associations among representations of events, and in such a way that Stimulus-Response connections would be simply a special case of the more general concept, perhaps clearly exemplified only in some forms of conditioning for human beings and in learning of lower organisms. (p. 7)

Like Stimulus Sampling Theory, adaptive networks provide a framework for theory development that builds cumulatively on the associative concepts originally established for simpler forms of learning. By extending assumptions regarding the representation of events, and the nature of the stimulus-response connections to be modified, the distributed network model integrates the powerful learning rule from the Rescorla–Wagner conditioning model with the stimulus-representation assumptions from Stimulus Sampling Theory. Each of these models has had a long and successful history within learning theory, the former being applied primarily to animal learning data and the latter being most often associated with human learning phenomena. The preliminary explorations reported here suggest that, within the formalisms of adaptive network theory, some of the shortcomings of each of these earlier models might be addressed by the strengths of the other.

ACKNOWLEDGMENTS

For their thoughtful comments and advice on this work I am indebted to Gordon Bower, W. K. Estes, Michael Hee, Irene Laudeman, Douglas Medin, Robert Nosofsky, Richard Shiffrin, and Robert Solso. This research was supported by Grant N00014-83K-0238 from the Office of Naval Research.

REFERENCES

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Atkinson, R. C. (1958). A Markov model for discrimination learning. *Psychometrika, 23,* 308–322.

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *Introduction to mathematical learning theory*. New York: Wiley.

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

Binder, A., & Estes, W. K. (1966). Transfer of response in visual recognition situations as a function of frequency variables. *Psychological Monographs: General and Applied, 80*(23), 1–26.

Binder, A., & Feldman, S. E. (1960). The effects of experimentally controlled experience upon recognition responses. *Psychology Monograph, 74*(496).

Binder, A., & Taylor, D. (1969). Effects of frequency and novelty in transfer. *Journal of Experimental Psychology, 80,* 91–94.

Blough, D. S. (1975). Steady-state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 104,* 3–21.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In E. Martin & A. Melton (Eds.), *Coding theory and memory*. Washington, DC: V. H. Winston.

Bower, G. H., & Hilgard, E. R. (1981). *Theories of learning*. New Jersey: Prentice-Hall.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America, 83*(4), 1615–1626.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145–154.

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. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press.

Estes, W. K. (1982). *Models of learning, memory, and choice*. New York: Praeger.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*(4), 556–571.

Estes, W. K., & Hopkins, B. L. (1961). Acquisition and transfer in pattern versus component discrimination learning. *Journal of Experimental Psychology, 61,* 322–328.

Feldman, J. A. (1985). Connectionist models and their applications: Introduction. *Special Issue of Cognitive Science, 9,* 1.

Flagg, S. F., & Medin, D. L. (1973). Constant irrelevant cues and stimulus generalization in monkeys. *Journal of Comparative and Physiological Psychology, 85*(2), 339–345.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2*(1), 50–55.

Gluck, M. A., & Bower, G. H. (1986). Conditioning and categorization: Some common effects of informational variables in animal and human learning. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society,* Amherst, MA. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gluck, M. A., & Bower, G. H. (1988a). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*(3), 225–244.

Gluck, M. A., & Bower, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

Gluck, M. A., & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General, 119*(1), 105–109.

Gluck, M. A., Bower, G. H., & Hee, M. R. (1989). A configural-cue network model of animal and human associative learning. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society.* Ann Arbor, MI. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hanson, S. J. (1990). A stochastic version of the delta rule. *Physica D, 42,* 265–272.

Hilgard, E. R., & Bower, G. H. (1975). *Theories of learning.* New Jersey: Prentice-Hall.

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

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

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

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.

LaBerge, D. L. (1961). Generalization gradients in a discrimination situation. *Journal of Experimental Psychology, 62,* 88–94.

LaBerge, D. L. (1962). A recruitment theory of simple behavior. *Psychometrika, 27*(4), 375–396.

Lashley, K. S. (1929). *Brain mechanisms and intelligence.* Chicago: University of Chicago Press.

Levine, M. (1970). Human discrimination learning: The subset sampling assumption. *Psychological Bulletin, 74,* 397–404.

Lovejoy, E. (1968). *Attention in discrimination learning.* San Francisco: Holden-Day.

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

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. (1976). Theories of discrimination learning and learning set. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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., & Robbins, D. (1971). Effects of frequency on transfer performance after successive discrimination training. *Journal of Experimental Psychology, 87*(3), 434–436.

Neimark, E. D., & Estes, W. K. (1967). *Stimulus Sampling Theory.* San Francisco, CA: Holden-Day.

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

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

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. (1976). Stimulus generalization: Some predictions from a model of Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 2,* 88–96.

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

Restle, F. (1957). Theory of selective learning with probability reinforcements. *Psychological Review, 64,* 182–191.

Robbins, D. (1970). Stimulus selection in human discrimination learning and transfer. *Journal of Experimental Psychology, 84,* 282–290.

Rudy, J. W., & Wagner, A. R. (1975). In W. K. Estes (Ed.), *Handbook of Learning and Memory, Vol. 2.* Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition* (*Vol. 1: Foundations*). Cambridge, MA: MIT Press.

Sejnowski, T. J. (1988). Neural populations revealed. *Nature, 332,* 308.

Shanks, D. R. (1989). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology.*

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

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

Suppes, P., & Atkinson, R. C. (1960). *Markov learning models for multiperson interactions.* New York: Stanford University Press.

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

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

Thompson, R. F. (1965). The neural basis of stimulus generalization. In D. J. Mostofsky (Ed.), *Stimulus generalization* (pp. 154–178). Stanford, CA: Stanford University Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review, Monograph Supplement,* 2(8), 28–31.

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.*) 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. 301–36). New York: Academic Press.

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