Approximately as appeared in: *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and J. Moore, Eds., pp. 497–537. MIT Press, 1990.

# Chapter 12

# Time-Derivative Models of Pavlovian Reinforcement

*Richard S. Sutton*
*Andrew G. Barto*

---

This chapter presents a model of classical conditioning called the *temporal-difference (TD) model*. The TD model was originally developed as a neuron-like unit for use in adaptive networks (Sutton and Barto 1987; Sutton 1984; Barto, Sutton and Anderson 1983). In this paper, however, we analyze it from the point of view of animal learning theory. Our intended audience is both animal learning researchers interested in computational theories of behavior and machine learning researchers interested in how their learning algorithms relate to, and may be constrained by, animal learning studies. For an exposition of the TD model from an engineering point of view, see Chapter 13 of this volume.

We focus on what we see as the primary theoretical contribution to animal learning theory of the TD and related models: the hypothesis that reinforcement in classical conditioning is the *time derivative* of a composite association combining innate (US) and acquired (CS) associations. We call models based on some variant of this hypothesis *time-derivative models*, examples of which are the models by Klopf (1988), Sutton and Barto (1981a), Moore et al (1986), Hawkins and Kandel (1984), Gelperin, Hopfield and Tank (1985), Tesauro (1987), and Kosko (1986); we examine several of these models in relation to the TD model. We also briefly explore relationships with animal learning theories of reinforcement, including Mowrer's drive-induction theory (Mowrer 1960) and the Rescorla-Wagner model (Rescorla and Wagner 1972).

Although the Rescorla-Wagner model is not a time-derivative model, it plays a central role in our exposition because it is well-known and successful both as an animal learning model and as an adaptive-network learning

algorithm (Widrow and Hoff 1960). We use the Rescorla-Wagner model as an example throughout the paper, and we use its limitations to motivate time-derivative theories of reinforcement. We also show that all predictions of the Rescorla-Wagner model can be obtained from a simple time-derivative theory of reinforcement closely related to that advocated by Mowrer and others in the 1950's.

One reason adaptive network models are of interest as animal learning theories is that they make predictions about the effect on learning of intra-trial temporal relationships. These relationships strongly influence learning, but little of modern animal learning theory deals explicitly with them. [1] The most well-studied effect is that of the CS–US inter-stimulus interval (ISI) on the effectiveness of conditioning. The attempt to reproduce the main features of this effect in a real-time computational model has driven much of the theoretical development of these models (e.g., Blazis and Moore 1987; Desmond, this volume; Grossberg and Levine 1987; Wagner 1981). In this paper, we systematically analyze the ISI behavior of time-derivative models, using realistic stimulus durations and both forward and backward CS–US intervals. The models' behaviors are compared with the empirical data for rabbit eyeblink (nictitating membrane) conditioning. We find that our earlier time-derivative model (Sutton and Barto 1981a) has significant problems reproducing features of these data, and we briefly explore partial solutions in subsequent time-derivative models proposed by Moore et al. (1986), Klopf (1988), and Gelperin et al. (1985).

The TD model was designed to eliminate these problems by relying on a slightly more complex time-derivative theory of reinforcement. In this paper, we motivate and explain this theory from the point of view of animal learning theory, and show that the TD model solves the ISI problems and other problems with simpler time-derivative models. Finally, we demonstrate the TD model's behavior in a range of conditioning paradigms including conditioned inhibition, primacy effects (Egger and Miller 1962), facilitation of remote associations, and second-order conditioning.

## *Theoretical Framework*

This section introduces the framework within which we discuss theories of reinforcement in classical conditioning. The presentation is largely tutorial, and readers already familiar with theories of classical conditioning may prefer simply to consult Equations 1 and 2 and then to skip to the following section, where we discuss time-derivative theories.

In a typical classical conditioning experiment, two stimuli, called the conditioned stimulus (CS) and the unconditioned stimulus (US), are paired in close succession. After sufficient pairings, the CS comes to produce a

response, called the CR, similar to the response originally produced only to the US. For example, in rabbit eyeblink conditioning, the CS might be the sound of a buzzer and the US might be a puff of air to the rabbit's eye that reflexively causes the eye to blink. After appropriate CS–US (buzzer–airpuff) pairings, the buzzer alone comes to elicit eyeblink CRs, providing evidence for the existence of a CS–US association. We consider experiments in which multiple CSs are used, either on the same or on different trials, but usually consider only one US. One theoretical interpretation of classical conditioning is that it is the process by which the animal infers causal relationships between stimulus events (Dickinson 1980). A related interpretation, to which we return later, is that classical conditioning is a manifestation of the animal's attempt to *predict* the US from cues provided by CSs.

A learning theory should predict how the associations between CSs and USs change. The most basic observation is that, in order for an association to change, the CS and US, or processes directly related to them, must occur at roughly the same time. Accordingly, almost all theories propose a multiplicative relationship between CS processing and contemporaneous US processing in determining the change in the association's strength, $V$:

$$\Delta V = (\text{level of US Processing}) \times (\text{level of CS Processing}). \qquad (1)$$

The amount of learning is thereby proportional to *both* the level of CS processing and the level of US processing, as long as these occur at the same time (cf., Dickinson 1980, p. 124; Wagner 1978). Some theories emphasize the effect of variations in CS processing, others the effect of variations in US processing. By virtue of the multiplicative interaction between these processes, many experimental results can be explained by reference to either.

For example, suppose two CSs, A and B, are presented simultaneously and paired with the US (written AB–US). It is generally found that one of the two CSs, say B, will be *overshadowed* by the other, A, in that it becomes much less strongly associated with the US than it would have in simple B–US pairings without A. Mackintosh's (1975) theory explains the deficit as due to competition between A and B for a limited CS processor; when presented together, one or both of the stimuli must get a significantly smaller share of the processor than it would if presented alone. Rescorla and Wagner's (1972) theory, on the other hand, explains the deficit by reference to competition for US processing. They propose that the level of US processing depends on how unexpected the US is. As A and B become associated with the US, they each reduce its unexpectedness, and thereby subtract from the amount of US processing available for the other. Again, at least one CS suffers a significant deficit.

As another example, consider a *blocking* experiment, in which extended A–US pretraining is followed by a second stage of AB–US training. The resulting association to B is found to be much weaker than that formed by an equivalent amount of AB–US training without A having been pretrained. According to Mackintosh's theory, A's pretraining identifies it as a useful CS; more attention is then paid to it, at B's expense, in second stage training. According to Rescorla and Wagner's theory, pretraining with A reduces the unexpectedness of the US in AB–US training, thus reducing the learning to both A and B during this stage.

*Reinforcement and Eligibility*

The US process is widely associated with the concept of Pavlovian reinforcement. Throughout this paper, we use the term *reinforcement* as a shorthand for "level of US processing" in the sense of Equation 1. It is also convenient to have a simple term for the level of CS processing. CS processing is associated with concepts such as attention, salience, stimulus traces, and rehearsal. Collectively, these concepts have to do with determining *which* CSs have their associations changed by reinforcement and which do not. We say that they determine which associations are *eligible* for change, should reinforcement occur; the level of processing of a CS is termed its *eligibility* (cf. Klopf 1972, 1982).

Using the new terms, we rewrite Equation 1 as

$$\Delta V = \text{Reinforcement} \times \text{Eligibility}. \tag{2}$$

Although the eligibility term is always positive (or zero), the reinforcement term can be either positive or negative. We refer to a positive reinforcement term as *positive reinforcement*, and to a negative reinforcement term as *negative reinforcement*. Because eligibility is always non-negative, increments in associative strength are always caused by positive reinforcement and decrements in associative strength are always caused by negative reinforcement.

In these terms, Rescorla and Wagner's theory explains blocking and overshadowing by reference to a theory of reinforcement, while Mackintosh's theory explains it by reference to a theory of eligibility. In this paper, we consider real-time models of both reinforcement and eligibility, but focus on reinforcement models. The models of eligibility we do consider—various forms of stimulus traces—are very simple compared to Mackintosh's theory or to other theories of CS processing (e.g., Pearce and Hall 1980; Lovejoy 1968; Zeaman and House 1963; Sutherland and Mackintosh 1971; Grossberg and Levine 1987). Like the Rescorla-Wagner model, our time-derivative models will most often invoke variations in reinforcement, i.e., in US processing, to explain results such as blocking and overshadowing.

*Trial-level Theories and Real-time Theories*

We have discussed one major distinction between theories—that some emphasize variations in reinforcement and others variations in eligibility. Another important distinction is whether their update equations, such as Equations 1 or 2, apply at every moment in time, both within and between trials, or only at the end of entire trials treated as wholes. Models that treat entire trials as wholes are called *trial-level* models. The Rescorla-Wagner model, for example, is a trial-level model: it makes predictions about what is learned from a trial based only on what CSs and USs were presented during the trial. Its predictions do not depend on the temporal relationships between these stimuli.

Models that apply continuously, on a moment by moment basis, are called *real-time* models. Hull's stimulus trace hypothesis (Hull 1939) is a simple example of a real-time model. According to that model, the internal representation of a CS persists for several seconds after CS offset. This *stimulus trace* determines the CS's eligibility, and thus the amount by which the CS's association is changed by reinforcement. The more time that passes between CS and US, the more the trace has faded by the time of the US, and thus the smaller the predicted change in associative strength. Unlike trial-level theories, such a real-time theory is able to make predictions about the effect on learning of *intratrial* temporal relationships among stimuli.

Both trial-level and real-time models have a long history in animal learning theory. Real-time models have the advantage that do not require a division of the animal's experience into trials by an experimenter or theorist. Trial-level models may *describe* animal learning behavior without specifying *how* it could come about. Such theories are harder to map into neural hardware, harder to convert into useful engineering algorithms, and ultimately less satisfying as scientific explanations. In addition, trial-level models do not consider intratrial temporal relationships, yet these are known to have significant effects on learning. Trial-level models can be applied successfully only when intratrial factors are held constant.

Nevertheless, the success of trial-level theories in predicting the results of experiments with constant intratrial temporal relationships is impressive. The trial-level Rescorla-Wagner model, for example, is the most influential current theory of classical conditioning. It has attained this status by accurately predicting the effects of a wide range of experimental manipulations while being a simple model clearly expressable by a few equations (as we discuss below). The challenge to real-time models is to achieve a comparable level of simplicity, clarity, and predictive accuracy while including variations in intratrial temporal relationships. Here we analyze one class of real-time models, which we call time-derivative models, relate them to previous models, and evaluate how well they have met this challenge.

## *Time-Derivative Theories of Reinforcement*

In this section we briefly present the Rescorla-Wagner model and explore its inherent limitations as a trial-level model. As a way of overcoming these limitations, we introduce the simple time-derivative theory of reinforcement used in the SB and DR models (Sutton and Barto 1981a; Klopf 1988). We show that this time-derivative theory makes all the same predictions as the Rescorla-Wagner model when the Rescorla-Wagner model applies, but, in addition, correctly accounts for phenomena beyond the scope of that model.

### *The Rescorla-Wagner Model*

The central idea of the Rescorla-Wagner model (Rescorla and Wagner 1972) is that learning occurs whenever events violate expectations, in particular, whenever the actual US level received on a trial differs from the level expected. In other words, Rescorla and Wagner hypothesized that reinforcement is the discrepancy between expected and actual US events. They denoted this discrepancy $\lambda - \bar{V}$, where $\lambda$ represents the actual US level on the trial and $\bar{V}$ represents the expected or *predicted* level. The predicted level, $\bar{V}$, is a composite or total prediction depending on the associative strengths of all the CSs present on the trial. Typically, it is assumed to be simply the sum of those associative strengths. The symbol $\lambda$ represents the effectiveness of the US received on the trial; if the US is absent, $\lambda$ is zero, otherwise $\lambda$ is some positive number combining, in an unspecified way, the US's intensity, duration, and temporal relationship with the CSs. If training is continued with the same CSs, then their composite prediction, $\bar{V}$, should approach $\lambda$.

To write the Rescorla-Wagner model in the form of Equation 2, a notation is needed for indicating whether or not a CS is present on the trial. Let $X_i = 1$ mean that the $i^{\text{th}}$ CS, $\text{CS}_i$, is present on the trial, whereas $X_i = 0$ means that $\text{CS}_i$ is absent. Let $V_i$ denote the associative strength of $\text{CS}_i$. With this notation, the prediction $\bar{V}$ is written $\bar{V} = \sum_i V_i X_i$, and the Rescorla-Wagner model is written

$$\Delta V_i = \beta(\lambda - \bar{V}) \times \alpha_i X_i, \tag{3}$$

where $\beta$ and $\alpha_i$ are positive constants depending on the US and CS respectively. For example, $\alpha_i$ generally reflects $\text{CS}_i$'s salience. Equation 3 is in the form of Equation 2, where $\alpha_i X_i$ is the eligibility and $\lambda - \bar{V}$ is the reinforcement.[2] The final term, $X_i$, is usually omitted from the equation defining the Rescorla-Wagner model, and instead it is stated in words that the equation applies only to the associative strengths of CSs that are present on a trial. Because $X_i$ is 1 for CSs present on a trial and 0 for those not present, it is clear that Equation 3 represents this selective application of the usual equation.

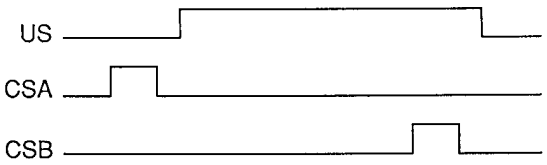


Figure 1
Illustration of variations in a US's reinforcing effect, $\lambda$, within a single trial. With a long duration US, a CS preceding its onset can become positively associated with the US, whereas a CS preceding its offset can become negatively associated.

A limitation of the Rescorla-Wagner model becomes apparent in second-order conditioning. In this procedure, a CS, A, is paired with the US, and then another CS, B, is paired with A. B can acquire a significant positive association with the US in this way (see Rescorla 1980a), a result contrary to the prediction of the Rescorla-Wagner model. On all trials on which B is present, the US does not occur, and thus $\lambda$ is zero, and the reinforcement $\lambda - \bar{V}$ in Equation 3 is negative or zero. Thus, the Rescorla-Wagner model incorrectly predicts that B's associative strength could only decrease or remain the same as a result of second-order conditioning.

To apply the Rescorla-Wagner model successfully to second-order conditioning one must hypothesize that pretrained CSs such as A create a positive $\lambda$. In particular, this $\lambda$ might be assumed proportional to A's associative strength. Once this has been done, the Rescorla-Wagner model correctly predicts the development of B's associative strength and the effect of substituting other CSs for B. The Rescorla Wagner model makes explicit predictions given a particular $\lambda$, but it does not specify what value $\lambda$ should have.

This limitation is particularly significant for a real-time theory of reinforcement, because $\lambda$ appears to vary even *within* trials. For example, Segundo et al. (1961) paired a long-duration shock US with two CSs, one preceding US onset, the other preceding US offset, as shown in figure 1. The CS that preceded US *onset* was found to develop a positive association with the US, while the CS that preceded US *offset* developed a negative association. Apparently, USs produce reinforcement—nonzero $\lambda$s—of opposite sign at their onset and offset (Mackintosh 1974, p. 113). The idea that *changes* in US level determine reinforcement, rather than the level itself, is the basis of time-derivitive theories of reinforcement.

### *The $\dot{V}$ Theory*

Second-order conditioning shows that CSs as well as USs can generate reinforcement if the CSs are associated with a US. Let us hypothesize that the reinforcing effects of a CS occur at its onset and offset, just as the

reinforcing effects of a US appear to occur at its onset and offset. In particular, suppose that each $CS_i$ with association of strength $V_i$ produces reinforcement $+V_i$ at its onset and $-V_i$ at its offset. The US can be viewed as having a large fixed association of strength $V_{US}$ such that it produces reinforcement $+V_{US}$ at its onset and $-V_{US}$ at its offset.[3] Thus, all stimuli, CSs and US, generate reinforcement $+V$ at their onset and $-V$ at their offset. For any given time, let $Y$ represent the sum of the associative strengths of all stimuli, including the US, that are present at that time. Note that $Y$ is not constant over a trial but changes as stimuli are presented and removed. Let $\dot{Y}$ denote the change in $Y$ over a small increment of time:

$$\dot{Y}(t) = Y(t) - Y(t - \Delta t).$$

Clearly, $\dot{Y}$ is zero except when some stimulus with a nonzero associative strength turns on or off, at which time it is $+V$ for an onset or $-V$ for an offset.[4] If several stimuli turn on or off simultaneously, then $\dot{Y}$ is the sum of all the individual reinforcements. Thus, we can formalize the central idea of the time-derivative theory as being that reinforcement at any time is given by $\dot{Y}$, the time-derivative of the net association, innate and acquired, between the current set of stimuli and the response. We call this the $\dot{Y}$ ("Y dot") theory of Pavlovian reinforcement.

The $\dot{Y}$ theory is sufficient to account for all the predictions of the Rescorla-Wagner model. Suppose the CSs present on a trial have simultaneous onsets and offsets, the offsets coinciding with US onset, and consider the reinforcement generated during the trial. The onsets of the CSs produce some reinforcement, but because no CS precedes this reinforcement, no CS associative strength is affected by it. There is a much better temporal relationship between the CSs and the reinforcement produced at the time of their joint offset (and the US onset). The CS offsets produce reinforcement of net strength $-\bar{V}$ and the simultaneous US onset produces reinforcement of strength $+V_{US}$. The net reinforcement at this time is thus $V_{US} - \bar{V}$. If we identify $V_{US}$ with $\lambda$, then this reinforcement is identical to that used in the Rescorla-Wagner model. Hence, for this special case, the $\dot{Y}$ theory and the Rescorla-Wagner model predict exactly the same changes in associative strengths.

So far we have ignored the reinforcement produced at US offset. This is appropriate if the US is long enough that reinforcement at its offset is in a very poor temporal relationship with the CSs. A shorter US complicates the analysis but does not change the conclusion. US offset produces reinforcement of $-V_{US}$. Presumably, however, the presented CSs are less eligible at this time because some time has elapsed since their offset; let us say they are half as eligible. The change in associative strength at this time is then $-\frac{1}{2}V_{US}$, for an overall change on the trial of

$$V_{\mathrm{US}} - \bar{V} - \frac{1}{2}V_{\mathrm{US}} = \frac{1}{2}V_{\mathrm{US}} - \bar{V},$$

which is again of the same form as the reinforcement term in the Rescorla-Wagner model, in this case with $\lambda = \frac{1}{2}V_{\mathrm{US}}$. Similar adjustments must be made to the value of $\lambda$ to deal with trace intervals between CS offset and US onset, but again the agreement with the Rescorla-Wagner model is retained. In general, for any trial in which the CSs begin and end simultaneously, the $\dot{Y}$ theory and the Rescorla-Wagner model predict exactly the same changes in associative strengths.

The $\dot{Y}$ theory was first formulated as stated here in our 1981 real-time model (Sutton and Barto 1981a). Hull's drive-reduction theory (Hull 1943) was perhaps the first to emphasize the role of changes in giving rise to reinforcement. According to that theory, however, only US offset produced reinforcement, and that reinforcement was positive rather than negative as it is in the $\dot{Y}$ theory. Mowrer's drive-induction theory (Mowrer 1950) is closer to the $\dot{Y}$ theory in proposing that US onset produces positive reinforcement, but does not assign a reinforcing role to offsets. Later Mowrer (1960) proposed that both onsets and offsets of both CSs and USs were reinforcers. Although Mowrer did not express his ideas as compactly as we have here in the $\dot{Y}$ theory, the basic idea can nevertheless be seen in his 1960 book. It is interesting to note, therefore, that the predictions of the Rescorla-Wagner model follow as a consequence of this idea. Klopf (1988) was the first to point out the relationships between the $\dot{Y}$ theory, which is also used in his DR model, and Mowrer's work.

The $\dot{Y}$ theory of reinforcement is not just consistent with the Rescorla-Wagner model, it is also a real-time extension of that model. The $\dot{Y}$ theory can be applied to any experiment, not only to those with simultaneous CSs, and thus has a larger scope than the Rescorla-Wagner model and the potential to unify disparate results. Before this can be explored, however, we should be clearer about the informal ideas we have been using regarding good and poor temporal relationships. Because this is a question of eligibility, we now turn to formalizing a real-time theory of eligibility.

### *Real-Time Theories of Eligibility*

As noted earlier, a theory of eligibility can include the effects of attention, salience, generalization, contrast, stimulus traces, and other phenomena involving the representation of CSs and the eligibility of their associations for being affected by reinforcement. Although all of these are important phenomena, they are beyond the scope of this paper, and we do not attempt to include them. As theories of eligiblity we consider only several simple kinds of stimulus traces.

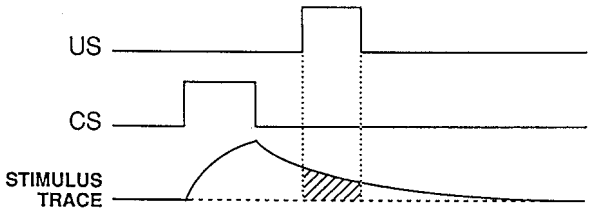


Figure 2
The role of the stimulus trace in bridging the trace interval between CS and US. The trace overlaps the US even though the CS does not.

*Hull's Stimulus Traces*

In classical conditioning experiments, the CS often terminates before the onset of the US. The interval of time between CS and US during which no stimuli are present is called the *trace interval.* Conditioning is found to be more effective as the trace interval decreases, but it can still occur at substantial intervals. Apparently, a CS leaves behind some short-term memory, or trace, indicating that is has recently been presented. Figure 2 illustrates the role of such a *stimulus trace* within a trial. The idea is to preserve the framework of Equations 1 and 2, in which contiguity of US and CS processes is required for learning to occur. Because the CS itself does not persist until the time of the US, some CS process must be postulated that does.

The time course marked CS in figure 2 represents the external, experimenter-defined stimulus. More important for learning, however, is the subject's internal representation of the stimulus. That these two can be different should be clear; for example, consider a brief flash of light and its afterimage. It is possible, then, that while the external CS terminates abruptly as shown in figure 2, the internal representation follows a different time course, perhaps one more like that shown for the stimulus trace. Hull (1939) proposed exactly this, that the stimulus trace is identical with the internal representation of the stimulus, and that all such internal representations persist for several seconds after the removal of the external stimulus.

*Eligibility Traces*

An alternative to Hull's stimulus trace is a trace that is distinct from the internal representation of the stimulus used to generate behavior. Such a distinct stimulus trace is responsible only for enabling learning, not for generating behavior; its effect is solely to influence the eligibility term of Equation 2. We distinguish this kind of stimulus trace from Hull's by calling it an *eligibility trace* (Klopf 1972, 1982). Of course, both kinds of stimulus traces could be used together.

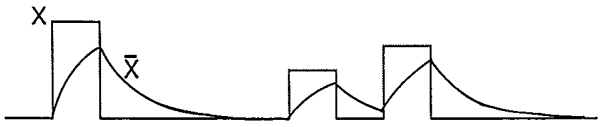


Figure 3
A simple eligibility trace. The time course of the trace $\bar{X}$ follows and lags behind the internal representation $X$ of the CS.

An advantage of using eligibility traces is that it is then not required that the internal representation of the CS be delayed or spread out in time. A delayed or spread out CS representation can make it difficult to produce rapid or precisely-timed responses. With eligibility traces, the internal representation is less constrained and can better support this sort of behavior. We have further discussed the advantages of eligibility traces elsewhere (Sutton and Barto 1981a; Barto and Sutton 1982). All the models we consider here use eligibility traces.

The simplest eligibility trace builds up while a CS is presented and fades away when it is removed, as illustrated in figure 3. Let $X_i$ denote the level of the internal representation of $\mathrm{CS}_i$ at each moment in time. For the moment, we assume that the internal representation is simply identical to the external one, that is, we assume that $X_i = 1$ when $\mathrm{CS}_i$ is present, and that $X_i = 0$ when $\mathrm{CS}_i$ is absent. The eligibility trace we denote by $\bar{X}_i$, and we think of it as a *running average* of recent values of $X_i$. The eligibility trace illustrated in figure 3 is obtained by continuously incrementing $\bar{X}_i$ at a fixed rate toward $X_i$. A complete specification for $\bar{X}$ is given in the appendix.

*The SB Model*

Our 1981 model, which Moore et al. (1986) called the Sutton-Barto, or SB, model, is obtained by combining the $\dot{Y}$ theory of reinforcement with the $\bar{X}$ eligibility trace:

$$\Delta V_i = \beta \dot{Y} \times \alpha_i \bar{X}_i,$$

where $\beta$ and $\alpha_i$ are positive constants as in the Rescorla-Wagner model. Since this is a real-time model, the equation applies at every moment within and between trials, rather than only once for each trial as in the Rescorla-Wagner model. We have previously shown the SB model to be consistent with a wide variety of empirical results, including all those where the Rescorla-Wagner model is applied, plus others including second-order conditioning, limited ISI dependency, and primacy effects (Sutton and Barto 1981a; Barto and Sutton 1982).

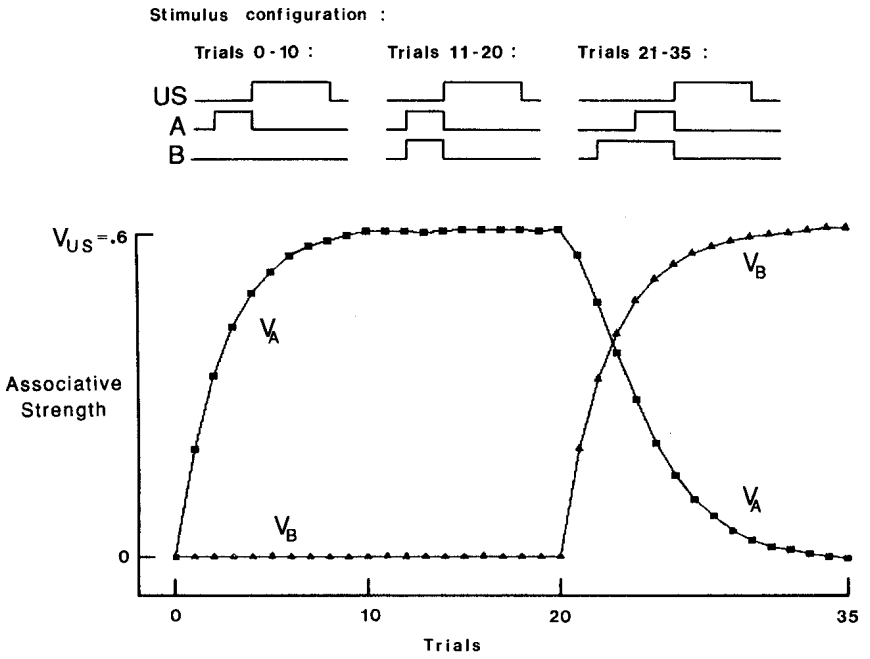


Figure 4
Illustration of the SB model. Shown are the associative strengths after each trial of a simulated experiment involving simple acquisition (Trials 1-10), blocking (Trials 11-20) and primacy (Trials 21-35). See text. Reprinted with notational changes from Sutton and Barto (1981a).

Figure 4 shows an example of the SB model's behavior. For the first 10 trials, a CS, A, is presented alone followed by the US; acquisition of the corresponding associative strength, $V_A$ , is shown. Trials 11-20 correspond to the second stage of a blocking experiment: A has already been conditioned, and now B is introduced with the same time course as A, followed by the US. The model shows complete blocking, with $V_B$ remaining at its intial value of zero during these trials. Finally, in trials 21-35, B is extended so that its onset precedes A. B is now in a good temporal relationship to pick up the positive reinforcement ( $\dot{Y}$ ) produced at the onset of A. Not only does B acquire associative strength during these trials, no longer being blocked by A, but A actually loses associative strength. Although we did not realize it at the time these results were first published, the SB model's prediction that A will lose associative strength under these conditions is novel and surprising. Why should a well-trained CS that continues to be paired with the US in a good temporal relationship lose associative strength just because a new CS is introduced with no initial association and in a

poorer temporal relationship to the US? One might expect the original CS to block or limit conditioning to the new CS, but the SB model predicts that the original CS rather than the new one will show a decrement in associative strength. Recently, Kehoe, Schreurs, and Graham (1987) tested and confirmed the prediction that the original CS can lose associative strength under these conditions. They also noted that alternative theories do not make this prediction and have considerable difficulty in explaining the result.

## *Problems with Time-Derivative Models*

Although the SB model successfully accounts for primacy effects, stimulus-context effects, and some effects of intratrial temporal relationships, it has also been found to have several problems. In this section, we review these problems and several new models that have been proposed to remedy them. In order to simplify the presentation we focus on two ways of evaluating a model. One is by comparison with empirical data regarding the effect on conditioning of the CS–US inter-stimulus interval. The second is by repeatedly presenting the model with a long serial-compound stimulus containing a different component CS for every time step before, during, and after the US. The response topography learned under these conditions is completely under the model's control and reveals something essential about the model.

### *Inter-Stimulus-Interval Dependency*

One of the main reasons for exploring real-time models is that they are able to make predictions based on intratrial temporal relationships among stimuli. One of the simplest cases in which this issue arises is that in which there is exactly one CS and one US. Empirically, the most important determinant of conditioning rate and asymptotic level is the time interval between the onset of the CS and the onset of the US, called the *inter-stimulus interval* or ISI. figure 5 shows the empirical relationship between the ISI and the effectiveness of two kinds of conditioning of rabbit nictitating membrane response. The shape of the empirical ISI dependency is roughly as shown here for all species and response systems, but the time course varies substantially (see, e.g., Macintosh 1974). The two kinds of conditioning for which data are shown are delay conditioning and what we call *fixed-CS conditioning* (see figure 6). In fixed-CS conditioning, the CS duration is fixed and independent of ISI. Fixed-CS conditioning includes trace conditioning, in which the ISI is greater than the CS duration, but also includes shorter and backward intervals. In delay conditioning, the CS duration is equal to the ISI, which is always positive.

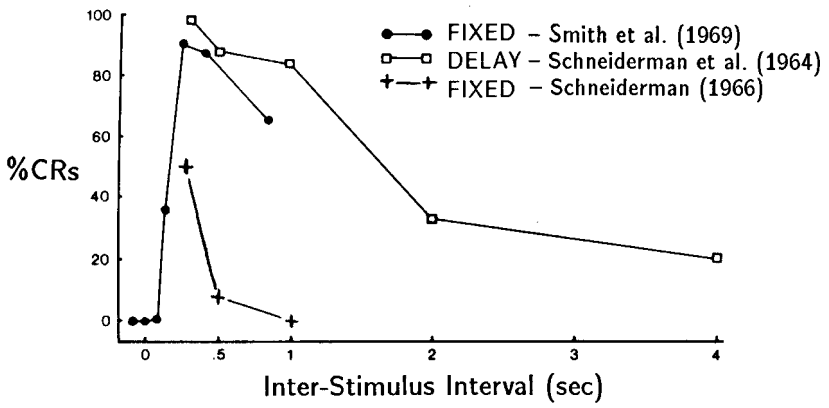


Figure 5
The empirical ISI dependency for the rabbit nictitating membrane response. Data is shown for both fixed-CS and delay conditioning (figure 6). The general shape of the ISI dependency is constant accross species and response systems, but its time course varies substantially.

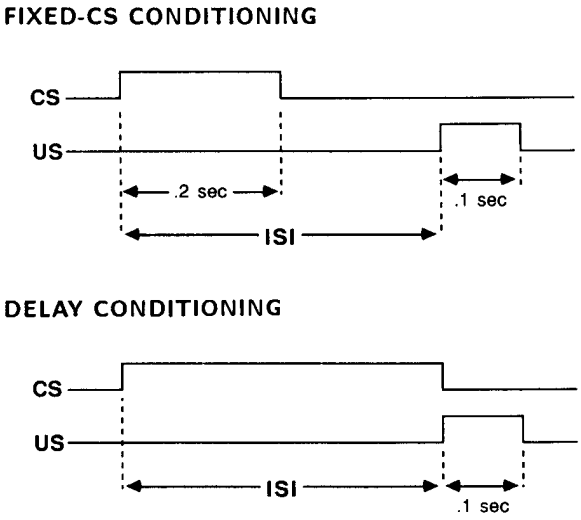


Figure 6
Temporal relationships in fixed-CS and delay conditioning. The indicated stimulus durations are commonly used in rabbit NMR conditioning. These durations were also used to obtain the simulation data shown in figures 8, 12, and 18, under the interpretation that one simulation time step is equivalent to 50 ms.

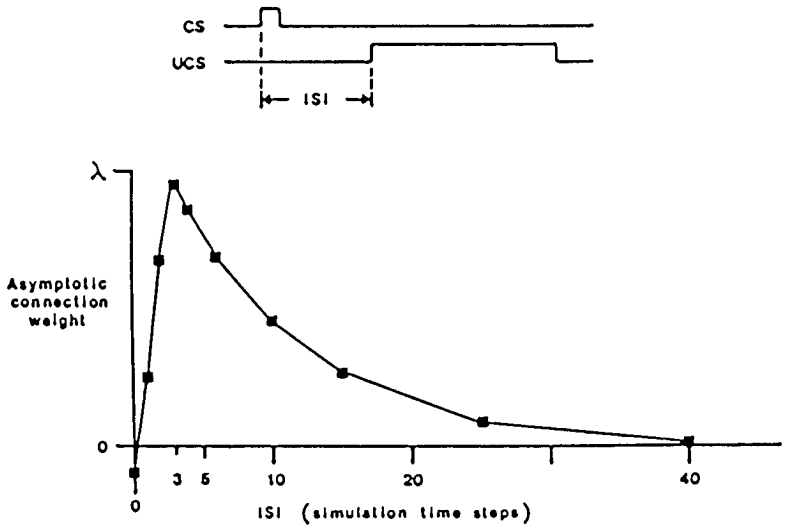


Figure 7
ISI dependency of the SB model for fixed-CS conditioning with a long US. (Reprinted from Sutton and Barto 1981a)

In Sutton and Barto 1981a, we compared the empirical data in figure 5 with the simulation data in figure 7 for fixed-CS conditioning of the SB model. In both cases, associative strength is near zero at zero ISI (simultaneous CS–US presentation), rises quickly to a maximum at intermediate ISIs, and then falls off gradually at long ISIs. If we identify each simulation time step with approximately 50 ms., then there appears to be a good match between model and data. However, this comparison is limited by the fact that the simulation used a very long US, equivalent to about 1500 ms., whereas a US of 100 ms. is more typical in real experiments. In addition, delay conditioning and backward fixed-CS conditioning paradigms were not simulated. If we repeat the simulation experiment, extended and made more realistic in these ways, we obtain the data shown in figure 8.

The SB model's ISI behavior shown in figure 8 deviates from the empirical data in figure 5 for delay conditioning at long ISIs and for fixed-CS conditioning at short forward and backward ISIs. In delay conditioning, the SB model predicts effective conditioning at all long ISIs, whereas the most prominent feature of the empirical ISI dependency of delay conditioning is the reduction in effectiveness of conditioning with increasing ISI. In fixed-CS conditioning, the SB model predicts strong inhibitory conditioning for both forward and backward conditioning when the CS and US overlap. The empirical data are not as clear here, as special tests must be run to detect

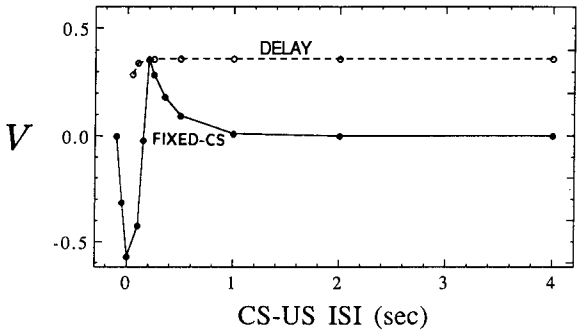


Figure 8
ISI dependency of the SB model. Shown is the CS associative strength $V$ after 80 acquisition trials as a function of the CS–US ISI. This simulation used a short-duration US. The intratrial temporal relationships were as shown in figure 6.

inhibitory associations, but the studies that have been done do not support the SB model's prediction of strong inhibitory conditioning with short USs (e.g., Prokasy et al. 1962). First we consider efforts to solve the delay conditioning problem and then efforts to solve the fixed-CS conditioning problem.

*Solving the Delay Conditioning Problem*

In Sutton and Barto 1981a, we explained the empirical reduction in effectiveness of delay conditioning at long ISIs by appealing to differences between external CS representations and internal (subjective) CS representations. Whereas the external CS remains constant during the ISI in delay conditioning, it is likely that the CS as perceived by the subject changes during the ISI. In particular, the beginning of the CS is probably represented more saliently than its end. For example, a long external CS such as that shown in figure 9a might give rise to a shorter internal CS representation such as that shown in figure 9b. If this were the case with the long CSs used in delay conditioning, then the SB model's ISI dependency for delay conditioning would look more like that for fixed-CS conditioning; that is, it would diminish with increasing ISI in qualitative accord with the empirical data.

The principal virtue of this explanation is that it leaves the SB model intact. The principal weakness of this explanation is that special internal CS representations have had to be hypothesized to deal with one of the simplest of classical conditioning experiments. As we consider more complex experiments, will the model's explanations involve increasingly complex hypotheses about internal represention? Relying on such hypotheses would

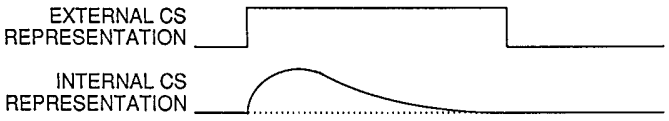


Figure 9
External and possible internal stimulus representations for a long overt CS.

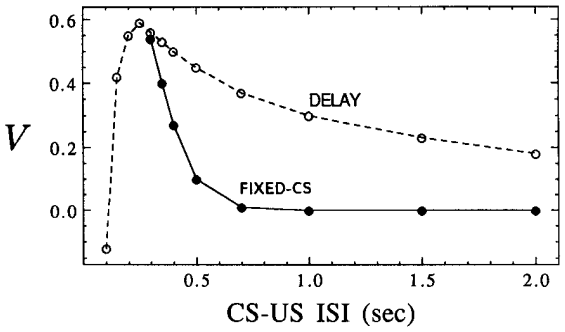


Figure 10
ISI dependency of the SBD model. The US duration was 30 ms; the CS duration in fixed-CS conditioning was 250 ms. These data are from Moore et al. (1986).

make it extremely difficult to unambiguously determine the predictions of the model.

Moore, Desmond, Blazis, et al. (Moore et al. 1986; Blazis et al. 1986), proposed modifying and extending the SB model to form the Sutton-Barto-Desmond (SBD) model. Although they were primarily concerned with matching behavioral and neurophysiological data on CR topography, their changes also resulted in a better match to the empirical ISI data for delay conditioning. They distinguished internal and external stimulus representations, but proposed a specific way of transforming one to the other so that this step could not be manipulated in an ad hoc manner. Among other changes, they hypothesized that the rate of decay of the eligibility trace increases as a function of CS duration. Figure 10 shows their simulation data for the ISI dependency of the SBD model in fixed-CS and delay conditioning. The SBD model correctly predicts a reduction in the effectiveness of delay conditioning at long ISIs. However, they also found weak inhibitory conditioning at some small ISIs in forward delay conditioning, depending on parameter settings (Blazis and Moore 1987). Data were not published for backward or simultaneous conditioning, but the model apparently did not significantly ease the SB model's problem of strong inhibitory conditioning (Blazis, personal communication).

Klopf (1988, 1986) proposed a simpler way of modifying the SB model to obtain weakened delay conditioning at long ISIs. In his Drive-Reinforcement (DR) model, eligibility does not increase during a CS but is triggered by CS onset and then follows a fixed time course whether or not the CS continues. Figure 11 illustrates such an *onset-triggered* eligibility for the case in which the time course of eligibility is a simple decay.[5] Because CS duration does not affect the time course of eligibility, fixed-CS and delay conditioning both lose effectiveness at long ISIs. Klopf (1988) demonstrated this in simulations, but has not published the complete ISI dependency of his model. Figure 12 shows the ISI dependency of what might be considered a *simplified DR model*—a model formed by using $\dot{Y}$ for reinforcement and onset-triggered eligibility whose time course is a simple delay.

Figure 12a shows the ISI dependency of the simplified DR model for fixed-CS conditioning after 80 trials. Note that strong inhibitory conditioning still occurs for simultaneous and backward CS–US presentation. Figure 12b shows the ISI dependency of the simplified DR model for delay conditioning after 80, 400, 2000, and 10,000 trials. In all cases, delay conditioning decreases in effectiveness at longer ISIs, in accord with the empirical data. However, the ISI at which the decrease begins increases with the number of conditioning trials. In fact, if conditioning proceeded to asymptote, delay conditioning at all ISIs would equal a maximal value determined by the intensity and duration of the US. For delay conditioning, this model predicts that associative strength increases toward the same high value for all ISIs, increasing faster at some ISIs than at others. Few experiments with animals have included the many thousands of trials that would be required to test this prediction, but the conventional interpretation of the available empirical results is that asymptotic conditioning level as well as rate of conditioning decreases at long ISIs (e.g., Bitterman 1964).

A second problem with onset-triggered eligibility is that it predicts that long CSs will not extinguish. Extinction in models using $\dot{Y}$ as reinforcement is normally caused by the decrement in $Y$, and hence negative $\dot{Y}$, at CS offset. However, if eligibility begins fading at the onset of a long CS, then it can be very small or zero by CS offset (see figure 11). Thus, it is incorrectly predicted that a sufficiently long excitatory CS will not extinguish through non-reinforced presentation, where "sufficiently long" is defined as longer than than the maximum ISI at which delay conditioning is effective. In general, the model predicts an inverse relationship between CS length and rate of extinction. The empirical data currently available do not directly contradict this prediction, but they are not supportive of it (e.g., see Schneiderman 1966). Morgan and Klopf (personal communication) have verified with simulations that the DR model makes these predictions.

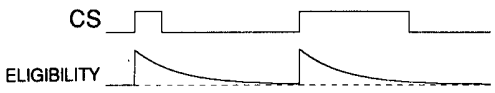


Figure 11
An onset-triggered eligibility trace. As in Klopf's DR model, the trace is incremented only at the onsets of CSs. For a very long CS, eligibility can nearly equal zero at CS offset.

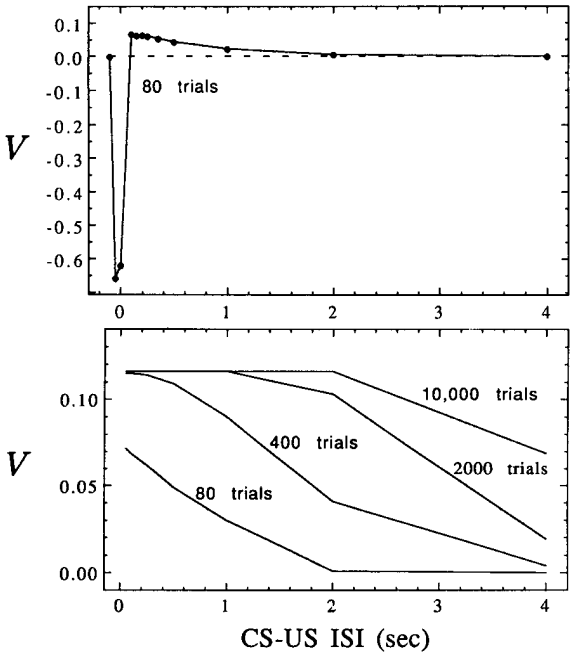


Figure 12
Interim and asymptotic ISI dependency of the simplified DR model. Top: Fixed-CS conditioning, 80 trials. Bottom: Delay conditioning, various numbers of trials. The parameter values used here were $\delta = 0.06$, $\beta = 1$ and $\alpha = 0.2$. The $\delta$ parameter was chosen to approximately match the effect of Klopf's (1988) choices for his learning rate parameters $c_1$ through $c_5$.

*Solving the Fixed-CS Conditioning Problem*

In fixed-CS conditioning, the SB model predicts strong inhibitory conditioning at simultaneous and near-simultaneous ISIs (figure 8), but this is not confirmed by the available empirical data. Strong inhibitory conditioning is predicted because of the good temporal relationship between the CS and the offset of the US. Inhibitory conditioning is in fact predicted by the SB model whenever the CS and US overlap, even in a forward arrangement with a long ISI. In the simplified DR model, inhibitory conditioning is predicted only for backward and simultaneous conditioning, but for those cases the inhibitory conditioning is very strong (see figure 12). The full DR model apparently makes similar predictions (Morgan and Klopf, personal communication), as does the SBD model (Blazis, personal communication). Empirically, backward and simultaneous conditioning have occasionally been found to produce weak inhibitory conditioning, but more often they produce weak *excitatory* conditioning (see Mackintosh 1974, 1983; Gormezano et al. 1983; Prokasy et al. 1962). Although further empirical studies are needed, it is clear that the predictions of strong inhibitory conditioning made by all of these models are counter to actual animal behavior.

One way of eliminating these problematic predictions is to use a modified $\dot{Y}$ theory in which only the *onsets* of USs create reinforcement, as in Gelperin, Hopfield and Tank's (1985) *Limax* model. This is effectively what we did in our original experiments with the SB model by using a very long US. If the US is long enough, its offset will occur when none of the CSs are eligible, and thus negative reinforcement at this time has little or no effect. This is the way we produced the fixed-CS ISI dependency for the SB model shown in figure 7, which shows less of a problem with inhibitory associations than does the ISI dependency shown in figure 8. However, ignoring US offset in this or any other way is questionable. For example, it is known that US duration affects conditioning (Gormezano, Kehoe and Marshall 1983, p. 233-4; Frey and Butler 1973) and that US offset can cause inhibitory conditioning (Segundo et al. 1961).

To better understand the SB model's problems when CS and US overlap, consider the case of complete overlap, that in which the CS and US begin simultaneously and end simultaneously. The reinforcement created at their joint onset causes no conditioning in the SB model because the CS is not yet eligible. However, the reinforcement at their joint offset, $-V_{\mathrm{US}} - V_{\mathrm{CS}}$ , is negative initially and causes $V_{\mathrm{CS}}$ to become negative. The learning process stabilizes when the reinforcement at offset is zero, that is, when $V_{\mathrm{CS}} = -V_{\mathrm{US}}$ . At this point the reinforcing effect of the US is exactly cancelled by that of the CS, both at offset and onset. This means that if a second CS is added that precedes the US, it will not acquire any associative strength. In fact, this prediction of the SB model does not depend on the simultaneous CS having been trained first. Even if the CS that precedes

the US is trained to a strong excitatory asymptote, a subsequently added simultaneous CS will become a strong inhibitor and cause the preceding CS to lose all its associative strength.

These are problematic predictions because conditionable stimuli with time courses similar to the time course of the US are invariably present in all conditioning experiments. For example, an airpuff to the eye acts as a US, but also produces conditionable stimuli. The time course of some of these stimuli will be similar to the time course of the US, whereas others will be longer, shorter, and with various delays; some may be initiated at US offset. We have shown that US cancellation results in $\dot{Y}$ models if *only* a US-simultaneous CS is present, but what if all these other CSs are present as well? This brings up the wider question of how the models behave when presented not with one or two stimuli, but with a whole collection of them. We now show that US cancellation tends to result in this case as well.

*Behavior in Response to Complete-Serial-Compound Stimuli*

The learning behavior of an animal or real-time model is often limited by the temporal pattern of CSs presented to it. For example, whenever no stimuli are present, $Y$ must be zero, and, whenever the CSs present are constant, $Y$ must be constant. Animals too are strongly influenced by the temporal pattern of CSs, but can partially overcome these limitations. For example, when animals are presented with a very long-duration CS, followed by the US, they eventually learn to repond differentially to the earlier and later portions of the CS. If the earlier and later portions are distinguishable, e.g., if they are tones of two different frequencies, then animals find it easier to respond differentially to them. Turning the original CS into a sequence of stimulus components, called a serial-compound stimulus, frees the animal to more easily exhibit what is in some sense its natural response. Taking this idea to its extreme, the animal or model could be presented with a distinguishable stimulus component for every small segment of time before, during, and after the US. If such a stimulus sequence completely covers the intratrial interval, then we call it a *complete* serial-compound stimulus, or CSC stimulus.

A *complete-serial-compound experiment* is an experiment in which a CSC stimulus is presented on each trial along with a US. In simulations of CSC experiments, a separate component CS is provided for *every time step* during a trial. The model is then able to produce a different, independent response level, $Y$, for each simulation time step during a trial. The behavior of $Y$ during a trial is not constrained by the CSs, but is entirely a function of the model's properties. It reveals what the model would do in every experiment with the US, if the model were not limited by the

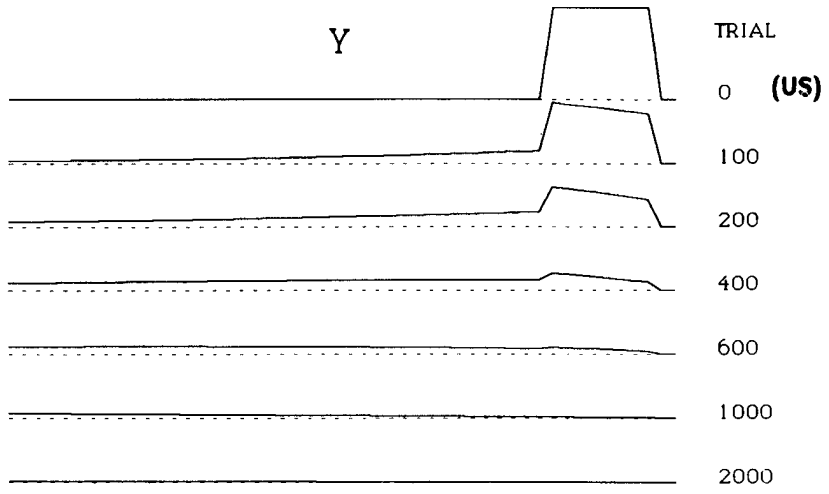


Figure 13
Intratrial behavior of $Y$ for the SB model in a CSC experiment. A separate CS has been provided for every time step before, during, and after the US. Shown in each graph is the behavior of the composite association $Y$ during a single trial. The first graph (trial 0) shows the timing of the US, as initially $Y$ is equal to the US signal. The height at each intratrial time shows the association to the component CS occurring at that time, plus the US association when the US is present. Initially, CSs preceeding the US become positively associated, but eventually the SB model learns only inhibitory associations such that all effects of the US are cancelled. The US is 8 time steps long, and the intratrial time period shown is 40 time steps long.

stimulus representation. Because $Y$ is presumably related to the CR, the intratrial behavior of $Y$ also has implications for theories of intratrial CR topography.

Figure 13 shows the intratrial behavior of $Y$ developing over trials in a CSC experiment with the SB model. The height at each intratrial time shows the association to the component CS occurring at that time (plus the US association for times when the US is present). Although there is some transient conditioning to CSs that precede the US, eventually this extinguishes, leaving only CSs occuring during the US with nonzero associative strengths. These CSs are conditioned inhibitors having identical strength, $-V_{\mathrm{US}}$ . Their effect is to exactly cancel the reinforcing effect of the US. The SB model thus predicts that there can be no asymptotic excitatory conditioning if a rich set of CSs are available.

The US cancellation problem shown most clearly by this CSC experiment is apparently an inherent consequence of the $\dot{Y}$ theory of reinforcement. For example, the same cancellation results if $\dot{Y}$ reinforcement is used in conjunction with an onset-triggered eligibility. Morgan and Klopf (personal correspondence) have verified that the DR model, which uses $\dot{Y}$

reinforcement, also cancels the effect of the US if presented with a simultaneous CS. This result also seems inevitable for the SBD model, which uses a modified $\dot{Y}$ theory of reinforcement. From the equations of the the SBD model it is clear at least that if the simultaneous CS initially cancels the US, then no reinforcement and thus no learning changes will occur. In general, learning stops in a $\dot{Y}$ theory of reinforcement whenever $\dot{Y} = 0$ is attained at all times during a trial. One way this can occur is by cancelling the US, but this is inappropriate for a model of classical conditioning.

One possible source of the US-cancellation problem is that primary and acquired reinforcers are treated nearly identically in the $\dot{Y}$ theory. They are identical except that the reinforcing effect of primary reinforcers is presumed to be fixed and permanent, whereas that of acquired reinforcers is subject to the learning process. This appeared initially to be consistent with the operational definition of primary reinforcers as reinforcers that retain their reinforcing effect even when repeatedly presented and not followed by another reinforcer, i.e., that do not extinguish. However, we have seen that the effects of primary reinforcers as well as the effects of acquired reinforcers tend to extinguish in models using $\dot{Y}$ reinforcement. This suggests that primary reinforcers should be modeled as being different from acquired reinforcers in some more essential way than is done in the $\dot{Y}$ theory.

We have discussed a number of attempts to solve problems with the ISI dependency of the SB model in delay and fixed-CS conditioning, none of which is completely successful. Most of the attempts to solve the problems in delay conditioning are modifications to the eligibility term. The fixed-CS conditioning problems, however, seem to implicate the reinforcement term. In the next section we take a more theoretical approach and derive a new reinforcement term, that used in the TD model, and show that it solves both kinds of problems.

### *The TD Model's Theory of Reinforcement*

Classical conditioning can be viewed as a manifestation of the subject's attempt to predict the arrival of the US. In terms of the Rescorla-Wagner model, $\bar{V}$ is the predicted US level on the trial and $\lambda$ is the actual US level. Their difference $\lambda - \bar{V}$ drives the learning process as the model's reinforcement term. How can we extend these trial-level ideas to form a real-time model? In this section we show how the *computational* theory of TD methods as prediction algorithms (Sutton 1988; see also chapter 13 of this volume) provides an answer to this question that solves the ISI problems of the other time-derivative models.

In the Rescorla-Wagner model, $\lambda$ for a trial depends on the US's intensity and duration. A more intense or longer US results in a larger $\lambda$ for that trial. In a real-time model, we let $\lambda$ change over time within a

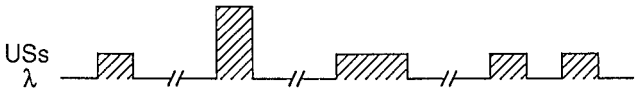


Figure 14
Time course of primary reinforcement, $\lambda$, for 4 USs of different intensities, durations, and repetitions. In all cases, the area under the curve represents the overall reinforcing effect.

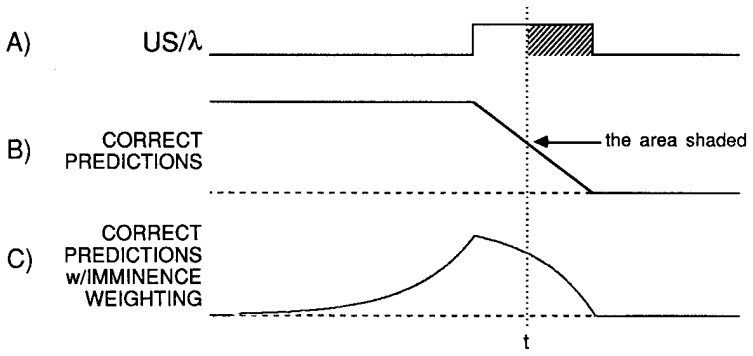


Figure 15
Time course of correct predictions near a US. A: Primary reinforcement; the correct prediction at each time $t$ relates to the future area under this curve. B: The correct predictions at each time $t$ of this future area. C: With imminence weighting, correct predictions fall as the area becomes temporally remote.

trial. Its value at each time represents the strength of the US, or rather the strength of its reinforcing effect, at that time. A more intense US is represented by a correspondingly larger $\lambda$ value at the times when the US is present, and a longer duration US is represented by a longer period of over which $\lambda$ is non-zero. A double US is represented by two intervals over which $\lambda$ is non-zero. In general, it is reasonable to propose that the area under the $\lambda$ curve (see figure 14) corresponds to the total primary reinforcement on the trial.

From a real-time perspective, then, we might consider the animal to be predicting the area under the $\lambda$ curve. Of course, at each time we would only be concerned with predicting the area for future $\lambda$'s. If the current time is half-way through a US, then the current prediction should be a prediction only of the remaining half, as shown in figure 15a. Figure 15b shows the correct predictions of future areas under the $\lambda$ curve at each point in time before, during, and after the presentation of the US. After the US, there is no future area and the correct prediction is zero. Before the US, all the area lies ahead and the correct prediction is constant and equal to the total area. During the US, the area remaining in the future

falls linearly from all of it at US onset to none at US offset. The arrow in figure 15b indicates the height that corresponds to the shaded area in figure 15a.

If the future areas in figure 15b are viewed as the predictions the subject is trying to learn, one problem is immediately apparent: the prediction level is equally high for all times prior to the US, whereas, empirically, animals seem to learn a weaker prediction for CSs presented far in advance of the US. The simple future-area view is also problematic theoretically. What if the US is so delayed that the experimenter considers it to be part of the next trial? Should the animal be predicting the sum of the areas of all the USs that will be delivered in the experiment? In its lifetime? Clearly, temporally remote primary reinforcement ( $\lambda$ values) should be discounted in some way. Primary reinforcement that is immediate should carry full weight; when slightly delayed, it should carry slightly less weight; when long-delayed, it should carry very little weight. In other words, upcoming primary reinforcement should be weighted according to its *imminence*. With imminence weighting, the correct prediction for each point in time near the US would look something like what is shown in figure 15c.

Another example of the effect of imminence weighting is shown in figure 16. Figure 16a shows a sequence of primary reinforcement created by a sequence of USs; this is a $\lambda$ curve. Figure 16b shows the *imminence weighting function*, specifying the way the weight given to primary reinforcement falls off with delay with respect to a particular time $t$. Figure 16c shows how the original sequence is transformed by the imminence weighting function applied at time $t$ to give reduced weight to delayed primary reinforcement. We propose that the quantity the subject is attempting to predict at time $t$ is the area under this curve rather than the area under the $\lambda$ curve. To obtain the correct predictions for other times, the weighting function is slid along the time axis so that its base starts at the time in question, the $\lambda$ sequence is reweighted according to the new position, and the new area is totalled. An example for another time $t'$ is shown in figures 16d and 16e. By repeating this process for every time, one obtains the sequence of correct predictions shown in figure 16f. This is what an animal should predict when faced with the US pattern in figure 16a, if it is attempting to predict imminence-weighted areas.

It is useful to formalize these ideas with explicit reference to time. For the moment we assume time is divided into small discrete steps. Let $\lambda_t$ denote the primary reinforcement received at time step $t$, where $t = 1, 2, 3, \ldots$ Let $\bar{V}_t$ denote the prediction made at time $t$ about values of $\lambda$ for times later than $t$. Our initial theory, without imminence weighting, is formalized by saying that the prediction should equal the sum of all future $\lambda$ values:

$$\bar{V}_t \text{ “}=\text{”} \lambda_{t+1} + \lambda_{t+2} + \lambda_{t+3} + \lambda_{t+4} + \cdots$$

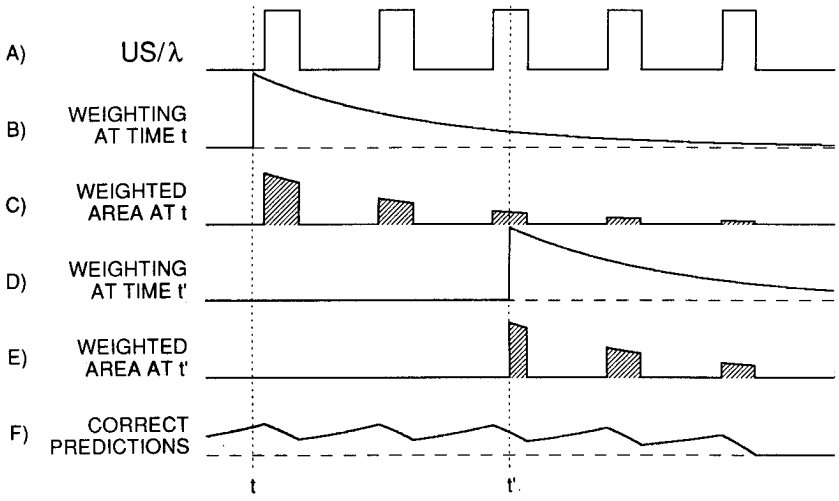


Figure 16
Imminence weighting. A: A temporal sequence of primary reinforcement (USs). B: Exponential imminence-weighting for time $t$ —the weight given at time $t$ to primary reinforcement at each later time. C: Primary reinforcement weighted for prediction at time $t$; the correct prediction at time $t$ is the area under this curve. D: Imminence weighting for time $t'$. E: Primary reinforcement weighted for prediction at time $t'$; the correct prediction at time $t'$ is the area under this curve. F: The correct predictions at each time; the heights at times $t$ and $t'$ equal the total areas in C and E.

where the quotation marks indicate that this is a desired relationship and not one that necessarily holds. We can introduce imminence weighting by discounting delayed primary reinforcement by some fraction $\gamma$, $0 \leq \gamma < 1$, for each step that it is delayed. One-step delayed primary reinforcement would then be discounted by $\gamma$, two-step delayed primary reinforcement by $\gamma^2$, three-step delayed by $\gamma^3$, and so on. The prediction at time $t$ should be

$$\bar{V}_t \text{ ``} = \text{''} \lambda_{t+1} + \gamma\lambda_{t+2} + \gamma^2\lambda_{t+3} + \gamma^3\lambda_{t+4} + \cdots \tag{4}$$

This is the form of discounting used to produce the desired predictions plotted in figures 15c and 16 (using a very small time step).

*Derivation of a Reinforcement Term*

If the goal is to obtain predictions as given by Equation 4, what can we conclude about the reinforcement term for use in our standard framework given by Equation 2? Sutton has recently developed a new computational theory of prediction methods, called *Temporal-Difference (TD) methods*, which suggests an answer. That theory provides a methodology for con-

structing TD learning methods specialized for predicting quantities in the form of Equation 4. We follow that methodology now to derive a suitable reinforcement term.

The discounted sum that we seek to predict, given by Equation 4, can be divided into two parts, one of which is the immediate reinforcement, and one of which is a new discounted sum containing all the later reinforcements:

$$\bar{V}_t \text{ “}=\text{”} \lambda_{t+1} + \gamma\Big[\lambda_{t+2} + \gamma\lambda_{t+3} + \gamma^2\lambda_{t+4} + \cdots\Big]. \tag{5}$$

The quantity in brackets is very similar to the overall sum to be predicted given by Equation 4. In fact, it is exactly what the prediction $\bar{V}_{t+1}$ is supposed to be. That is, if we write out Equation 4 for the desired value for $\bar{V}_{t+1}$:

$$\bar{V}_{t+1} \text{ “}=\text{”} \lambda_{t+2} + \gamma\lambda_{t+3} + \gamma^2\lambda_{t+4} + \gamma^3\lambda_{t+5} + \cdots,$$

we see that we can exactly substitute $\bar{V}_{t+1}$ into Equation 5 to obtain

$$\bar{V}_t \text{ “}=\text{”} \lambda_{t+1} + \gamma\bar{V}_{t+1}.$$

Thus, we can simply state the desired prediction for one time step in terms of the primary reinforcement and desired prediction for the next time step. We want the prediction at each time step to equal the primary reinforcement received on the next step plus the next prediction (discounted by $\gamma$). The discrepancy or error is then the difference between these quantities:

$$\lambda_{t+1} + \gamma\bar{V}_{t+1} - \bar{V}_t.$$

This discrepancy is much like the discrepancy used in the Rescorla-Wagner model, $\lambda - \bar{V}$, where the role of $\lambda$ in their model is being taken here by $\lambda_{t+1} + \gamma\bar{V}_{t+1}$. This suggests that we use this discrepancy directly as a reinforcement term. If this is done in combination with the $\bar{X}$ model of eligibility, one obtains the temporal-difference (TD) model (Sutton and Barto 1987):

$$\Delta V_i = \beta\Big(\lambda_{t+1} + \gamma\bar{V}_{t+1} - \bar{V}_t\Big) \times \alpha_i\bar{X}_i,$$

where $\beta$ and $\alpha_i$ are positive constants as in the Rescorla-Wagner model, and where the equation applies at each moment in time as in all real-time models. [6]

Although the TD model does not use $\dot{Y}$ as its reinforcement, we consider it to be a time-derivative model of reinforcement. In the theory above, $\bar{V}_t$ and $\lambda_{t+1} + \gamma\bar{V}_{t+1}$ are viewed as predictions, formed on successive time steps, of the same quantity, a discounted sum of $\lambda$ values. The discrepancy between these two predictions is thus a discrete-time analog of the time derivative of the prediction of that quantity. This discrepancy rep-

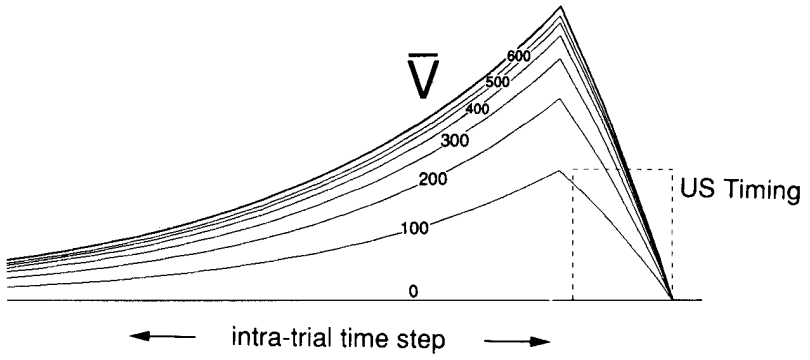


Figure 17
Intratrial behavior of $\bar{V}$ for the TD model in a CSC experiment. The highest curve represents the theoretically correct predictions as given by Equation 4. The lower curves are the predictions generated by the TD model in a CSC experiment after various numbers of trials, as indicated. A different component CS is presented for every time step during the trial. The height at any intratrial time represents the associative strength of the component CS presented at that time. The peak of the predictions is one time step before US onset—when the US is temporally closest but still lies entirely in the future. The US is 8 time steps long, and the intratrial time shown is 40 time steps.

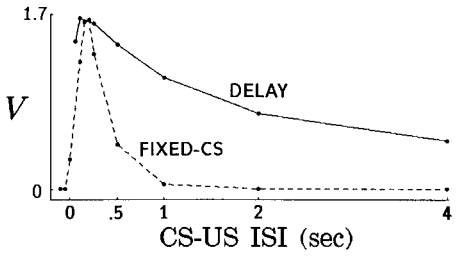


Figure 18
ISI dependency of the TD model. Unlike the other models, the TD model's ISI dependency is a good match to the empirical data in figure 5. These associative strengths were obtained after 80 trials. See figure 6 for the temporal relationships between stimuli.

resents how the model's expectation of discounted future $\lambda$ changes from one time step to the next. In Chapter 13, we show in more detail how this view of the TD model fits into a computational framework.

Since the TD model is based on a theory of reinforcement in which the correct predictions follow the time course shown in figure 15c, one might expect it to produce similar actual predictions in a complete-serial-compound experiment. Figure 17 shows that indeed it does. The highest curve represents the theoretically correct intratrial predictions as given by Equation 4. The lower curves are the prediction topographies generated by the TD model at various trials of a simulation experiment in which a different CS is provided for each time step before, during and after the US. The actual predictions gradually approach the ideal ones. This shows that the new model has solved the problems with cancellation of USs demonstrated earlier for the $\dot{Y}$ models.

Figure 18 shows the ISI dependency of the TD model for fixed-CS and delay conditioning. These curves are a good match to the empirical data for rabbit NMR in figure 5. Delay conditioning decreases in effectiveness at long ISIs, and there is no problem of strong inhibitory conditioning in fixed-CS conditioning at near-zero ISIs.

*Demonstrations of the TD model*

The TD model seems to solve many of the problems with other time-derivative models, but does the TD model retain the desirable properties of these models on the wide range of experimental conditions in which they have been explored? We have previously shown (Sutton and Barto 1987) that it does, and that, in fact, it is in slightly better accord with the data than is the SB model. In addition, we have shown that the TD model is consistent with the data on serial-compound experiments to a degree that has not been shown for previous models such as the SB model or Klopf's DR model.[7] Below we review some of these demonstrations. We present results showing the behavior of the TD model in a range of conditioning paradigms including blocking, facilitation of remote associations, primacy effects, and second-order conditioning.

The TD model exhibits complete blocking if first-stage training is conducted until asymptotic associative strength is achieved and the CS added in the second stage has exactly the same time course as the first CS. This follows directly from Equation 2 and the use of an $\bar{X}$ eligibility trace. From Equation 2, the only way to have a different change in associative strength for two CSs is for their eligibilities to differ. But the $\bar{X}$ eligibility traces of two CSs with the same time course are identical. Therefore, if the pretrained CS no longer undergoes any change in associative strength in the second stage of a blocking experiment, then neither can the new CS. The

pretrained CS remains fully associated, and the new CS remains with zero associative strength.

One of the well-known failings of the Rescorla-Wagner model is that, in its simplest form, it predicts that a CS with a negative association, a *conditioned inhibitor*, will extinguish if presented alone. Empirically, this extinction has not been observed (Zimmer-Hart and Rescorla 1974). However, this incorrect prediction results from the assumption that the composite association, $\bar{V}$, is a simple sum of associative strengths of the CSs present on a trial. If one assumes instead that $\bar{V}$ is the sum if that sum is positive, and zero otherwise, then the model correctly predicts that conditioned inhibitors will not extinguish. Donegan, Gluck and Thompson (1989) and others have noted this for the Rescorla-Wagner model; Moore et al. (1986) showed essentially the same thing for the SB model (by similarly assuming $Y$ is always non-negative), and Klopf (1988) has shown this for the DR model. We (Sutton and Barto 1987) followed their example and specified the TD model's $\bar{V}$ to be $\sum V_i X_i$ when that sum is positive and to be zero when the sum is negative. We now show that this produces the correct behavior in a conditioned inhibition experiment. [8]

Figure 19 shows the behavior of the TD model (with $\bar{V}$ restricted to be non-negative) in a conditioned inhibition (CI) training regime. In CI, reinforced and unreinforced trials of the two types shown in figure 19a are intermixed. $CS^+$ is followed by the US except in the presence of $CS^-$. $CS^+$ is found empirically to become positively conditioned whereas $CS^-$ becomes a conditioned inhibitor. This result was also found in the simulation. In the extinction phase of the simulated CI experiment, both stimuli were presented individually without the US. The result shown in figure 19 is the same as that found empirically: the association to the excitor extinguishes, but the association to the inhibitor does not (Zimmer-Hart and Rescorla 1974).

Real-time conditioning models are interesting primarily because they make predictions for a wide range of situations that cannot be represented by trial-level models. These situations involve conditionable stimuli that occur together but not strictly simultaneously. A compound stimulus whose components do not both begin and end at the same time is called a *serial compound*. It should be recognized that almost all learning involves serial compounds, either because the animal distinguishes earlier and later portions of a stimulus that may be viewed as a single stimulus by the experimenter, or because the animal's behavior gives rise to a predictable sequence of situations leading to reinforcement, as in maze running. Kehoe (1982) surveys the theoretical issues and empirical results relevant to serial-compound conditioning.

One of the theoretical issues arising in serial-compound conditioning concerns the facilitation of remote associations. It has been found that

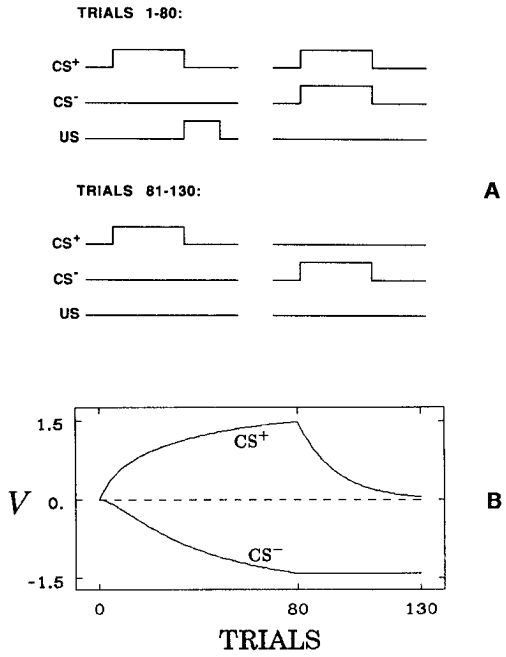


Figure 19
Conditioned inhibition and its extinction in the TD model. In this and in all following simulations, $\bar{V}_t$ was forced to be non-negative. A) Time traces showing the two kinds of trials presented alternately in a conditioned inhibition experiment (trials 1-80) and in a subsequent attempt to extinguish the resultant associations (trials 81-130). B) Behavior over trials of the associative strengths of $CS^+$ and $CS^-$. During acquisition, the associative strength of $CS^+$ becomes positive, while the associative strength of $CS^-$ becomes negative. The association of $CS^+$, but not of $CS^-$, is extinguished by nonreinforcement. Both CSs were 200 ms. in duration and the US was 100 ms. in duration.

if the empty trace interval between the CS and the US is filled with a second CS to form a serial compound stimulus, then conditioning to the first CS is facilitated. Figure 20b shows the behavior of the TD model in a simulation of such an experiment, the timing details of which are shown in figure 20a. Consistent with the experimental results (see Kehoe 1982), the model shows facilitation of both the rate of conditioning and the asymptotic level of conditioning of the first CS due to the presence of the second CS.

As discussed earlier, a strength of real-time models is their ability to make predictions about the effects on conditioning of intratrial temporal relationships. One of the best-known demonstrations of such an effect is an experiment by Egger and Miller (1962) that involves two overlapping CSs in a delay configuration as shown in figure 21a. Although CSB is in a better temporal relationship with the US, the presence of CSA reduces conditioning to CSB substantially as compared to controls in which CSA is absent. Figure 21b shows the same result being generated by the TD model in a simulation of this experiment.

Earlier we discussed similar results for the SB model (figure 4) from an earlier paper (Sutton and Barto 1981a). That simulation experiment differed from the Egger-Miller experiment in that the shorter CS was given prior training until it was fully associated with the US. When the longer, earlier CS was introduced, the association to the pretrained short CS decreased as training continued. As we discussed earlier, this is a surprising and then-untested prediction, subsequently confirmed by Kehoe, Schreurs, and Graham (1987), who also noted that alternative (non-time-derivative) theories do not make this prediction and have considerable difficulty in explaining the result. The behavior of the TD model under these conditions is shown in figure 22. This behavior is actually in slightly better accord with the data than is the SB model's behavior, in that the association to the pretrained short CS is reduced after the introduction of the long CS, but not completely eliminated.

Figure 23 shows the behavior of the TD model in a second-order conditioning experiment. In the first phase (not shown in the figure), CSB is pretrained with the US. In the second phase, CSA is paired with CSB in the absence of the US, in the sequential arangement shown in figure 23a. Empirically, CSA is found to acquire associative strength even though it is never paired with the US. In the TD model, CSA first acquires a substantial association and then that association and CSB's association extinguish. The same pattern is seen empirically.

Figure 24 shows the ISI dependency of the TD model for second-order conditioning. It plots the associative strength after 10 trials as a function of the CSA–CSB ISI. This ISI curve differs from the CS–US ISI curve shown in figure 18 in that here simultaneous presentation results in the forma-

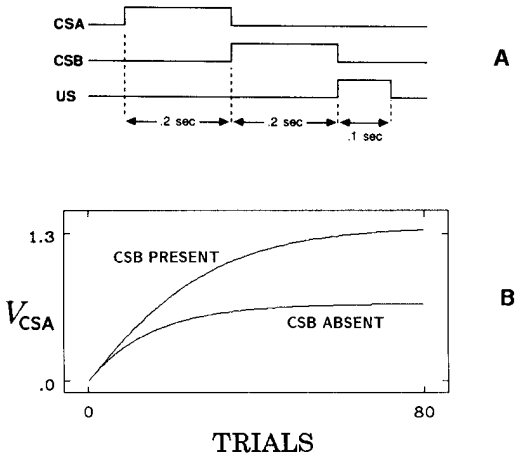


Figure 20
Facilitation of a remote association by an intervening stimulus in the TD model. A: Temporal relationships among stimuli within a trial. B: The behavior over trials of CSA's associative strength when CSA is presented in a serial compound, as in A, and when presented in an identical temporal relationship to the US, only without CSB.

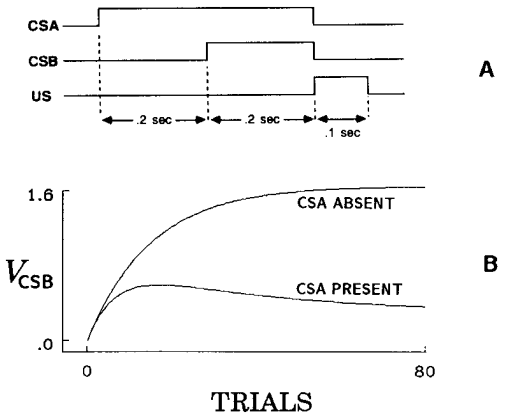


Figure 21
The Egger-Miller or primacy effect in the TD model. A: Temporal relationships among stimuli within a trial. B: The behavior over trials of CSB's associative strength when CSB is presented with and without CSA.

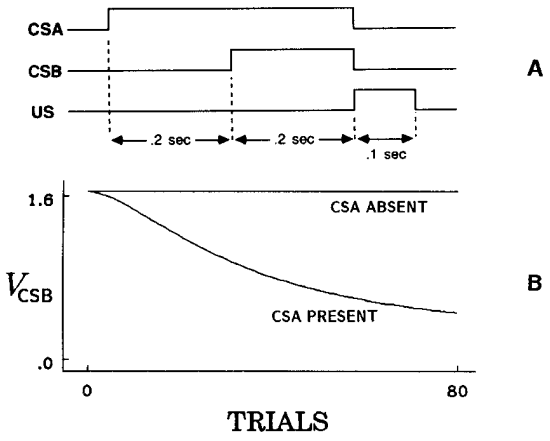


Figure 22
Temporal primacy overriding blocking in the TD model. A: Temporal relationships between stimuli. B: The behavior over trials of CSB's associative strength when CSB is presented with and without CSA. The only difference between this simulation and that shown in figure 21 was that here CSB started out fully conditioned—CSB's associative strength was initially set to 1.653, the final level reached when CSB was presented alone for 80 trials, as in the "CSA-absent" case in figure 21.

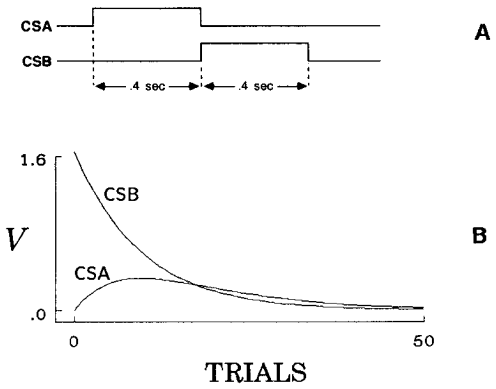


Figure 23
Second-order conditioning of the TD model. A: Temporal relationships between stimuli. B: The behavior of the associative strengths associated with CSA and CSB over trials. The second stimulus, CSB, has an initial associative strength of 1.653 at the beginning of the simulation.

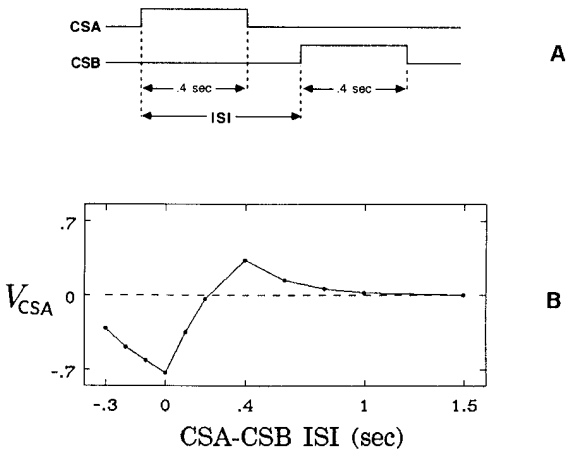


Figure 24
Effect of the CSA-CSB ISI on second-order conditioning of TD model. A: Temporal relationships between stimuli. B: Resultant value of CSA's associative strength after 10 trials as a function of CSA-CSB ISI.

tion of a large negative association instead of a small positive one. The TD model treats the reinforcement due to USs and previously conditioned CSs differently: US signals directly cause reinforcement, whereas *changes* in the signals of previously conditioned CSs cause reinforcement. Thus, in simultaneous presentation, a US's reinforcement is delivered thoughout the presentation, whereas a previously conditioned CS delivers reinforcement only at its onset, and negative reinforcement at its offset, so that a simultaneously paired CS will be much more affected by the negative reinforcement than by the positive reinforcement.

Empirically, second-order conditioning is observed to occur with both simultaneous and sequential CSA–CSB pairings. To explain this observation in terms of the TD model we must appeal to indirect associations, which are outside the scope of the model per se. That is, the model clearly predicts that no direct CSA–US association will develop, but does not preclude the development of both CSA–CSB and CSB–US associations, which together could have the effect of a CSA–US association. This explanation of second-order conditioning is in fact partially confirmed empirically. One observed difference between simultaneous and sequential second-order conditioning is that responding to CSA is eliminated by extinguishing CSB after simultaneous second-order conditioning, but not after sequential second-order conditioning (Rescorla 1980b). This suggests that simultaneous second-order conditioning in fact does not result in a direct CSA–US association. These simulation results also suggest the prediction

that simultaneous pairing in second-order conditioning should result in a negative CSA–US association. To our knowledge, this has not been tested.

*Conclusion*

The hypothesis that Pavlovian reinforcement is the time derivative of a composite US and CS association accounts for many aspects of classical conditioning. As Mowrer noted thirty years ago, it provides a unified account of single-CS acquisition, higher-order conditioning, primacy effects, and many instrumental learning phenomena that we have not considered here. As we noted in 1981, it also accounts for a wide range of stimulus context effects by virtue of its reduction to the Rescorla-Wagner model for the special case of simultaneously presented CSs. Once formalized and combined with appropriate stimulus traces, time-deriviative models also predict the effects of variations in intratrial temporal relationships. In particular, we have shown that the TD model reproduces salient features of the empirical data in all three of these areas.

In comparing different time-derivative models, we have focussed on their predictions about the effect of the CS–US interstimulus interval on single-CS acquisition conditioning. The predictions of our 1981 model deviate from the empirical data for both fixed-CS and delay conditioning. The related models proposed by Moore et al. (1986) and Klopf (1988) ease some but not all of these problems. In particular, all of these models incorrectly predict that a CS simultaneous with the US will become strongly inhibitory and block conditioning to CSs that precede the US. Only the TD model reproduces the main features of the empirical ISI dependency without additional assumptions about subjective stimulus representations.

A distinguishing feature of the TD model is that it is based on a theory about the *function* of classical conditioning. It is based on the supposition that the goal of learning is to accurately predict at each point in time the imminence-weighted sum of future US intensity levels. Given this goal, the equations of the TD model follow from a computational theory of adaptive prediction algorithms (Sutton 1988; Barto, Sutton and Watkins, this volume). The TD model thus both predicts features of classical conditioning behavior and provides an account of their function as part of a mechanism for accurate prediction.

Finally, we note that the TD model is of course not a *complete* model of classical conditioning. Among the major classes of phenomena not directly addressed by the model are attention, salience, configuration, and learning to learn (e.g., see models by Mackintosh 1975; Pearce and Hall 1980; Kehoe 1988; Gelperin, Hopfield and Tank; Schlimmer and Granger 1986; Grossberg and Levine 1987). Models of many of these phenomena could be added to the TD model as a "front end", or pre-processing stage,

intervening between the external stimuli and their representation to the TD model. At the output end, the TD model has significant implications for CR topography, but would need to be augmented with a reponse rule before forming a full model of response generation (e.g., see models by Moore et al. 1986; Frey and Sears 1978; Blazis and Moore 1987; Desmond, this volume). Finally, we have noted that indirect associations, as revealed, e.g., in sensory preconditioning experiments, are beyond the immediate scope of the model. To include indirect associations, the model would need to be extended from an adaptive-element model to a adaptive-*network* model (e.g., see models by Moore and Stickney 1980; Schmajuk and Moore 1986; Sutton and Barto 1981b; Sutton and Pinette 1985).

*Acknowledgments*

The authors acknowledge their indebtedness to John Moore, Harry Klopf, and Jim Kehoe; this work is the result of a long collaboration with them. We also thank Jim Morgan, Diana Blazis, Mark Gluck, and Chuck Anderson for helpful discussions, Marcy Rosenfield for figure reconstructions, and particularly Jim Morgan for performing simulation experiments to verify our analyses of the DR model. Richard Sutton thanks the enlightened management of GTE Laboratories for making this work possible. Andrew Barto acknowledges the support of the Air Force Office of Scientific Research through grant AFOSR-87-0030.

*Appendix: Details of TD Model Simulations*

The equations in the text were left slightly ambiguous in order to avoid a distracting complication of the notation. For clarity, the equations actually used in the TD model simulations are given here in full, with explicit reference to time:

$$V_i(t+1) = V_i(t) + \beta \left( \lambda(t+1) + \gamma \left\lfloor \sum_j V_j(t) X_j(t+1) \right\rfloor - \left\lfloor \sum_j V_j(t) X_j(t) \right\rfloor \right) \alpha \bar{X}_i(t+1),$$

where $\lfloor x \rfloor$ is $x$ unless $x < 0$, in which case it is $0$, and

$$\bar{X}_i(t+1) = \bar{X}_i(t) + \delta \Big( X_i(t) - \bar{X}_i(t) \Big).$$

When a stimulus was present, the corresponding input signal ( $X_i(t)$ or $\lambda(t)$ ) was set to 1, and when the stimulus was absent, the signal was set to 0. All associative strengths, $V_i$ , and eligibility traces, $\bar{X}_i$ , were zero at the start of training, except in the few cases explicitly noted.

The time interval between trials was long enough for all traces to fall to zero. Since no stimuli were presented during the inter-trial interval, it is clear that reinforcement will be zero during this time, and that therefore no learning will occur. Thus, the inter-trial interval was simulated simply by setting the traces to zero.

The parameter values used were $\alpha = 0.1$, $\beta = 1.0$, $\delta = 0.2$, and $\gamma = 0.95$. These values were chosen to approximately match ISI data for the rabbit nictitating membrane response (figure 5) under the interpretation that each time step corresponds to 50 ms. To produce the same behavior under a different interpretation of the time step, different parameter values must be used. For example, if one switched to an interpretation of a simulation time step as 10 ms., then five times as many time steps would have to occur in the same amount of clock time. Each of the learning rates $\alpha$ and $\delta$ would therefore have to be reduced by a fifth, to 0.02 and 0.04 respectively, so that approximately the same amount of learning would occur. The rate at which imminence-weighting decreases, determined by $\gamma$, must also be reduced by a fifth. This is done by cutting the drop from 1.0 to $\gamma$ by a fifth. In this case, by changing $\gamma$ from 0.95 to 0.99. Finally, note that the associative strengths, $V_i$, represent predictions of future areas, that is, of sums of future $\lambda$ values. Sampling time more finely means there will be proportionally more $\lambda$ values to add up in the same amount of clock time. This means that associative strengths learned using different time scales can only be compared if the time scale is taken into account. For example, figure 18 shows associative strengths under optimal conditions reaching values of approximately 1.7. If a five times smaller time step was used, then the associative strengths would instead reach approximately $5 \times 1.7 = 8.5$. All of these adjustment rules are only approximate, but should give good results as long as the time step is kept small.

*Notes*

1. Wagner's (1981) SOP model is a notable exception.
2. We drop the US-dependent constant $\beta$ in discussing reinforcement terms because we generally consider only a single US.
3. One may ask, "what is the US's association with?" The US can either be considered to be associated with itself, just as the CSs are associated with the US, or both US and CSs can be considered to be associated with the response produced by the US. In either case, it makes sense for the US's association to be large and permanent.
4. For this purpose we ignore changes in the individual associative strengths.
5. Klopf (1988) actually proposed a time course of eligibility that was inverted-U shaped, like the empirical ISI dependency. However, for our purposes this difference is probably not significant.
6. See the appendix for specification of the TD model with explicit reference to time and for a listing of the parameter values used in the simulation experiments that follow.
7. Klopf and Morgan (personal communication) have recently obtained results for the DR model that in some cases parallel those presented here.

8. The outcomes of the experiments previously described are not affected by this redefinition of $\bar{V}$, because none of them involved inhibitory associations.

## *References*


Barto, A.G., Sutton, R.S. (1982) Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research 4*: 221–235.

Barto, A.G., Sutton R.S., Anderson, C.W. (1983) Neuronlike elements that can solve difficult learning control problems. *IEEE Trans. on Systems, Man, and Cybernetics, SMC-13*, No. 5, 834–846.

Bitterman, M.E. (1964) Classical conditioning in the goldfish as a function of the CS–US interval. *Journal of Comparative and Physiological Psychology 58*: 359–366.

Blazis, D.E.J., Desmond, J.E., Moore, J.W., Berthier, N.E. (1986) Simulation of the classically conditioned nictitating response by a neuron-like adaptive element: A real-time variant of the Sutton-Barto model. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 176-186.

Blazis, D.E.J., Moore, J.W. (1987) Simulation of a classically conditioned response: Components of the input trace and a cerebellar neural network implementation of the Sutton-Barto-Desmond model. Technical Report 87-74, Computer and Information Science Dept., University of Massachusetts, Amherst.

Dickinson, A. (1980) *Contemporary Animal Learning Theory*. Cambridge University Press.

Donegan, N.H., Gluck, M.A., Thompson, R.F. (1989) Integrating behavioral and biological models of classical conditioning. In: *Computational Models of Learning in Simple Neural Systems (Volume 22 of the Psychology of Learning and Motivation)*, R.D. Hawkins and G.H. Bower, Eds. Academic Press.

Egger, D.M., Miller, N.E. (1962) Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology 64*: 97–104.

Frey, P.W., Butler, C.S. (1973) Rabbit eyelid conditioning as a function of unconditioned stimulus duration. *Journal of Comparative and Physiological Psychology 85*: 289–294.

Frey, P.W., Sears, R.J. (1978) Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review 85*: 321–348.

Gelperin, A., Hopfield, J.J., Tank, D.W. (1985) The logic of *Limax* learning. In: *Model Neural Networks and Behavior*, A. Selverston, Ed. Plenum Press.

Gormezano, I., Kehoe, E.J., Marshall, B.S. (1983) Twenty years of classical conditioning research with the rabbit. In: *Progress of Psychobiology and Physiological Psychology*, J.M. Sprague and A.N. Epstein, Eds., 198–274. Academic Press.

Grossberg, S., Levine, D.S. (1987) Neural dynamics of attentionally modulated Pavlovian conditioning: blocking, interstimulus interval, and secondary reinforcement. *Applied Optics 26*: 5015–5030.

Hawkins R.D., Kandel, E.R. (1984) Is there a cell-biological alphabet for simple forms of learning? *Psychological Review 91*: 375–391.

Hull, C.L. (1939) The problem of stimulus equivalence in behavior theory. *Psychological Review 46*: 9–30.

Hull, C.L. (1943) *Principles of Behavior*. Appleton-Century-Crofts.

Jacobs, R.A. (1988) Increased rates of convergence through learning rate adaptation. *Neural Networks 1*: 295–307.

Kehoe, E.J. (1982) Conditioning with serial compound stimuli: Theoretical and empirical issues. *Experimental Animal Behavior 1*: 30–65.

Kehoe, E.J. (1988) A layered network model of associative learning: Learning to learn and configuration. *Psychological Review 95*: 411–433.

Kehoe, E.J., Schreurs, B.G., Graham, P. (1987) Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response. *Animal Learning and Behavior 15*: 455–464.

Klopf, A.H. (1972) Brain function and adaptive systems—A heterostatic theory. Air Force Cambridge Research Laboratories Special Report No. 133 (AFCRL-72-0164). Also DTIC Report AD 742259 available from the Defense Technical Information Center, Cameron Station, Alexandria, VA 22304.

Klopf, A.H. (1982) *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere.

Klopf, A.H. (1986) A drive-reinforcement model of single neuron function: An alternative to the Hebbian neuronal model. In J.S. Denker (Ed.) *Neural Networks for Computing* (Conference Proceedings *151*, American Institute of Physics, 265–270).

Klopf, A.H. (1988) A neuronal model of classical conditioning. *Psychobiology 16*: 85–125.

Kosko, B. (1986) Differential Hebbian Learning. In J.S. Denker (Ed.) *Neural Networks for Computing* (Conference Proceedings *151*, American Institute of Physics, 277–282).

Lovejoy, E. (1968) *Attention in Discrimination Learning*. Holden Day.

Mackintosh, N.J. (1974) *The Psychology of Animal Learning*. Academic Press.

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

Mackintosh, N.J. (1983) *Conditioning and Associative Learning*. Oxford University Press.

Moore, J.W., Desmond, J.E., Berthier, N.E., Blazis, D.E.J., Sutton, R.S., Barto, A.G. (1986) Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing and interstimulus intervals. *Behavioral Brain Research 21*: 143–154.

Moore, J.W., Stickney, K.J. (1980) Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. *Physiological Psychology 8*: 207–217.

Mowrer, O.H. (1950) *Learning Theory and Personality Dynamics*. Ronald Press.

Mowrer, O.H. (1960) *Learning Theory and Behavior*. Wiley. (Krieger Edition, 1973)

Pearce, J.M., Hall, G. (1980) A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review 87*: 532–552.

Prokasy, W.F., Hall, J.F., Fawcett, J.T. (1962) Adaptation, sensitization, forward and backward conditioning, and psuedo-conditioning of the GSR. *Psychol. Rep. 10*: 103–106.

Rescorla, R.A. (1980a) *Pavlovian Second-order Conditioning*. Erlbaum.

Rescorla, R.A. (1980b) Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavioral Processes 6*: 339–351.

Rescorla, R.A., Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical Conditioning II*, A.H. Black and W.F Prokasy, Eds., 64–99. Appleton-Century-Crofts.

Schlimmer, J.C., Granger, R.H. (1986) Simultaneous configural classical conditioning. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 141-153.

Schmajuk, N.A., Moore, J.W. (1986) A real-time attentional-associative network for classical conditioning of the rabbit's NMR. *Proceedings of the Eighth Annual Conf. of the Cognitive Science Society*, 794–807.

Schneiderman, N. (1966) Interstimulus interval function of the nictitating membrane response of the rabbit under delay and trace conditioning. *Journal of Comparative and Physiological Psychology 62*: 397–402.

Schneiderman, N., Gormezano, I. (1964) Conditioning of the nictitating membrane of the rabbit as a function of the CS–US interval. *Journal of Comparative and Physiological Psychology 57*: 188–195.

Segundo, J.P., Galeano, C., Sommer-Smith, J.A., Roig, J.A. (1961) Behavioral and EEG effects of tones 'reinforced' by cessation of painful stimuli. In: *Brain Mechanisms and Learning*, J.F. Delafresnaye, Ed. Blackwells Scientific Publishing.

Smith, M.C., Coleman, S.R., Gormezano, I. (1969) Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous and forward CS–US intervals. *Journal of Comparative and Physiological Psychology 69*: 226–231.

Sutherland, N.S., Mackintosh, N.J. (1971) *Mechanisms of Animal Discrimination Learning*. Academic Press.

Sutton, R.S. (1984) Temporal credit assignment in reinforcement learning. Ph.D. dissertation, University of Massachusetts.

Sutton, R.S. (1988) Learning to predict by the methods of temporal differences. *Machine Learning 3*: 9–44.

Sutton, R.S., Barto, A.G. (1981a) Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review 88*: 135–171.

Sutton, R.S., Barto, A.G. (1981b) An adaptive network that constructs and uses an internal model of its environment. *Cognition and Brain Theory Quarterly 4*: 217–246.

Sutton, R.S., Barto, A.G. (1987) A temporal-difference model of classical conditioning. *Proceedings of the Ninth Conference of the Cognitive Science Society*, 355–378.

Sutton, R.S., Pinette, B. (1985) The learning of world models by connectionist networks. *Proceedings of the Seventh Annual Conf. of the Cognitive Science Society*, 54–64.

Tesauro, G. (1986) Simple neural models of classical conditioning. *Biological Cybernetics 55*: 187–200.

Wagner, A.R. (1978) Expectancies and the priming of STM. In: *Cognitive Processes in Animal Behavior*, S.H. Hulse, H. Fowler, and W.K. Honig, Eds. Erlbaum.

Wagner, A.R. (1981) SOP: A model of automatic memory processing in animal behavior. In: *Information Processing in Animals: Memory Mechanisms*, N.E. Spear and R.R. Miller, Eds., 5–48. Erlbaum.

Widrow B., Hoff, M.E. (1960) Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104.

Zeaman, D., and House, B.J. (1963) The role of attention in retardate discrimination learning. In: *Handbook of Mental Deficiency: Psychological Theory and Research*, N.R. Ellis, Ed., pp. 159-223. McGraw-Hill.

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