8 Models of Acquisition and Preference

Michael L. Commons,
Eric W. Bing,
Charla C. Griffy,
and Edward J. Trudeau
Harvard University

ABSTRACT

Previously, preference and acquisition studies have only been concerned with developing acquisition models based on observed behavior for known reinforcement parameters. Schedules of reinforcement were taken into account when models were constructed, and learning curves were developed as if the subject knew the reinforcement parameters. The learning parameters were based upon a response strengthening model derived from Thorndike's Law of Effect (Thorndike, 1898, 1911).

This study attempts to predict acquisition based on probabilistic values of reinforcers from the subject's limited perspective. We assert that underlying instrumental conditioning, two respondent conditioning steps take place. We also assert the effect of having a response predict the occurrence of a following reinforcer decreases as a hyperbolic fashion as the time between them increases. Thus, a model is constructed in which the learning curve changes over time as the subject's behavior progresses. A final predicted learning curve becomes apparent data point by data point, as each is processed by this model. The resulting predicted behavior for one such model is compared to the observed behavior, and an analysis is then carried out for accuracy of fit.

MODELS OF ACQUISITION AND PREFERENCE

Quantitative acquisition of preference studies with nonhumans essentially began in the early 1980's (e.g., Commons, Woodford, Boitano, Ducheny, & Peck, 1982; Herrnstein, 1982; Herrnstein & Vaughan, 1980; Myerson & Mizzen, 1980). In preference studies, organisms select among schedules of reinforcement by

either responding on one operandum (key) more often than on others or responding on an operandum that selects one of the schedules. The reinforcement schedules deliver reinforcers (S^{R+}) to the organisms based upon some rule, for example, following the first response (R) after one minute has elapsed since the last reinforcer had been delivered (Fixed Interval, FI). In the late 1980s there have been a few such studies (Myerson & Hale, 1988; Vaughan & Herrnstein, 1987). Preference experiments, in which organisms make real choices and experience real outcomes form an even smaller subclass of such studies (e.g., Bailey, 1988; Bailey & Mazur, submitted; Commons, Woodford, Boitano, Ducheny, & Peck, 1982; Myerson & Mizzen, 1980). Corresponding theories and experimental results are so few that a fairly detailed history of them is possible.

Preference situations can be characterized as discrimination situations, in which the stimuli associated with the responses are quite easily distinguishable. For example, pecks on the red-left key (R_L) are reinforced on one schedule. Pecks on the green-right key (R_R) are reinforced on another schedule. An acquisition experiment studies how an organism changes its behavior in response to a changing set of stimuli. At the beginning of acquisition, one set of schedules has been in effect for a long time. Organisms show stable preferences reflected by relatively constant choice probabilities or rates. The schedules are then changed and the resulting changes in response frequency or rate as performance restabilizes constitute the acquisition data.

In the short history of data based quantitative preference studies, simple concurrent, schedules have been examined by Myerson (Myerson & Hale, 1988; Myerson & Mizzen, 1980). He placed pigeons on two concurrent schedules whose values (programmed reinforcement frequency) were periodically switched. In such a concurrent schedule, each response (R) that met the reinforcement contingency was reinforced (SR+). The clock for one schedule continued to run even though responding on the other schedule was taking place. Therefore, for time-based reinforcement schedules, responding on the left increased the likelihood that responding on the right would be reinforced.

Myerson and his associates proposed the Kinetic Model to explain their acquisition data. For example each of the schedules could be random ratio schedules.

$$R_L \rightarrow S^{R+} \text{ with } p = .1$$
$$R_R \rightarrow S^{R+} \text{ with } p = .2 \qquad (1)$$

The word ratio reflects that fact that there is a ratio of the number of responses, n, to the 1 response that is reinforced (Reynolds, 1968). With a random ratio schedule, a response is reinforced with probability p, a Poisson process (Schoenfeld & Cole, 1972).

At the same time, two other groups of researchers were examining acquisition of preference. First, (Herrnstein, 1982; Herrnstein & Vaughan, 1980; Vaughan & Herrnstein, 1987) set forth their notion of melioration. Herrnstein had previously established his matching law as a model describing the relationship between

responding and reinforcement after responding has stabilized. Melioration is a model describing what an organism will do if reinforcement conditions are changed. Vaughan (1981, 1985) made clear that the "value" of the reinforcement schedules being preferred was carried by the stimuli that preceded the responses. Otherwise, response rate would not have momentarily dropped when key color was changed but programmed reinforcement rate stayed constant. Another aspect of their theory is that the mechanism of change depended on the obtained rate of responding and obtained rate of reinforcement rather than the programmed rate of reinforcement. No specific experiments were analyzed at the time, and few details in their model were specified.

Second, Commons, Woodford, Boitano, Ducheny, and Peck (1982) examined the acquisition of preference by using a concurrent chain procedure as shown next:

Pattern of reinforced pecks across cycles, C_1				
	C_1	C_2	C_3	C_4
Completing requirement 1, VI-12 seconds for R_L, leads to:	R-0	R-S^{R+}	R-0	R-0
Completing requirement 2, VI-12 seconds for R_R, leads to:	R-0	R-S^{R+}	R-S^{R+}	R-0

(2)

In the concurrent-chain schedule they used, completing requirement 1 (a VI-12 seconds component) led to four trials (C_1 through C_4) described by the pattern of reinforced pecks. If a response was reinforced, it was delivered at the end of the cycle, C_4; no reinforcement is indicated by 0. The length of C_1 was 3 seconds. In a variable interval schedule (VI), after a variable length with a mean of t seconds of time after the last reinforcer, the first response is reinforced. Commons, Woodford, et al. were interested in relating the value of reinforcement obtained in a preference situation to the value obtained in the situation where reinforcement schedules were discriminated. In a discrimination situation, some form of responding indicates which schedule has been in effect previously. Correct indications are reinforced. Concurrent chains, in which completing one schedule leads to another, were adopted at the suggestion of Nevin (1978, personal communication). Nevin suggested that concurrent schedules were better understood than simple choice procedures for assessing preference. The Commons, Woodford, et al's model was midway between others in that the brief obtained rates of responding were used, but steady-state values of obtained reinforcement were also used. In any case, a titration procedure was used, in which reinforcer values were switched for choice outcome, and shifts toward stability were recorded.

We know of only three other data based studies. Dreyfus (1985, March; 1985, April) presented some data on what happens with concurrent schedules when values are switched. A final report of that data is forthcoming. Myerson and Hale (1988) also studied acquisition of preference using concurrent schedules. They

suggest that their data is inconsistent with a melioration model and consistent with their kinetic model.

Most recently Bailey and Mazur (Bailey, 1988; Bailey & Mazur, submitted) ran the simplest experiment. They first stabilized pigeons pecking on a simple two choice situation. They then examined a number of theories including melioration (Vaughan, 1981), probability learning (Estes, 1959) as well as the kinetic model.

Relating Instrumental Preference to Classical Conditioning Neural Networks

This paper presents a conceptual reduction model of how preference can be conceived of as two steps of respondent (classical) conditioning. It then presents pilot models integrating the reduction and the Linear Noise Model (Commons, Woodford, & Ducheny, 1982), a model on effect of the delay of reinforcement. These models are of a behavioral nature and do not compete with neural network models of the same processes (Grossberg, 1987). They may be useful, however, to network modeling of the acquisition of preference. One of the most important aspects of neural networks is that they are formed by not only modeling behavioral data but also modeling neural network processes. As most of the papers appearing

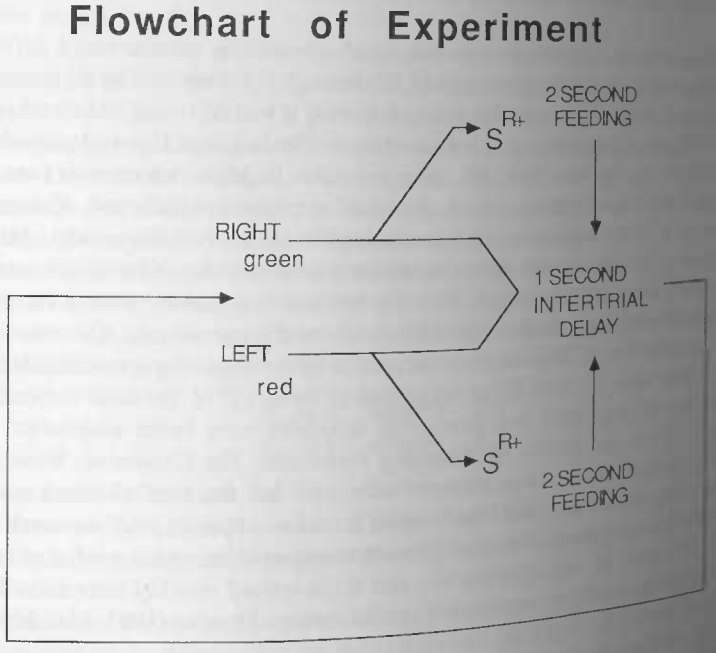


FIG. 8.1. Flowchart of experiment.

in this volume attest, models of conditioning generally refer to classical (respondent) conditioning rather than instrumental (operant) conditioning. Grossberg (1971a, 1971b, 1974, 1982a, 1982b, 1987) is the most prominent exception because he integrates operant and respondent conditioning and develops neural networks of them. There is a long history of relating these two forms of conditioning. Two factor theories suggest that operant and respondent condition are separate processes. Single factor theories suggest that although there are surface differences in the procedures, the underlying processes are the same. Our strategy is to show how operant conditioning is related to respondent conditioning and how they differ, so that the methods used for respondent conditioning may be applied to operant conditioning.

The Difference Between Operant and Respondent Conditioning

In order to model the acquisition of operant preference, one must draw upon the neural network studies of respondent (classical) condition. We identify and examine differences between the two to show why our reduction of operant to respondent conditioning is necessary in order to use respondent conditioning neural network results.

Operant conditioning and respondent conditioning cannot be immediately reconciled. First, in operant conditioning, an environmental stimulus (S) followed by an operant reinforcer never produces a response, whereas in respondent conditioning it does. Second, in classical conditioning there is no necessity to follow the conditioned response (CR) with an unconditioned stimulus US or operant reinforcer (S^{R+}). Third, simple classical conditioning between the environmental stimulus and the stimulus that elicits the operant response fails. The unobserved but inferred stimulus that elicits the operant response is an unconditioned stimulus (us). In classical conditioning, presentation of the environmental stimulus followed by an unconditioned stimulus leads to conditioning. In the operant case, presenting the environmental stimulus followed by the (us) that elicits the operant response leads to extinction of the operant response. Although the differences are not limited to these three, these distinctions are specifically addressed by reduction.

Although operant conditioning and respondent conditioning cannot be immediately reconciled, they can be united by reducing operant conditioning to respondent conditioning. For background on how we conceive of the contingencies in operant conditioning, the following reduction of operant to respondent conditioning is presented. Pilot behavioral models of operant preference acquisition are also introduced that update and combine features of the Commons and Woodford (Commons, Woodford, & Ducheny, 1982) model with features of the Herrnstein and Vaughan (1980) model. Together, the reduction and behavioral models provide a conceptual springboard, from which a neural network of preference acquisition might be built.

The Reduction of Operant Conditioning to Respondent Conditioning

It is argued here that the functions of operant reinforcement can be derived from the functions and properties of respondent pairing operations. Two important functions of operant reinforcement are the strengthening of responding and the establishment of discriminative control by events. We propose that the operant response is not spontaneously emitted behavior, but rather is an acquired response to certain appropriate stimuli.

The two extra requirements that preserve the explanatory power of two factor theories are the two respondent pairing steps in the present reduction: (a) the salience or "what to do" pairing step, in which the internal causal events (US) that precede the operant response (R/UR_1) become salient to the subject, who thereby learns "what to do" to receive reinforcement; and (b) the environmental control or "when to do" pairing step, in which the now salient internal stimulus-conditioned response complex (US-CR_2) is paired with the neutral environmental stimulus (S_1/NS). The subject thereby learns "when to do" the operant behavior and under what circumstances.

In respondent conditioning the first step is unnecessary, because the unconditioned stimulus (US) is already salient to the subject. Because of requirement (b), the organism cannot learn to make the operant response under different circumstances unless it has been exposed to various situations. More reinforced responding within one situation will not lead to generalization of learned behavior in other situations.

We posit the existence of an event of unknown origin within the brain called the internal unconditioned stimulus (*us*). This *us* is a behavioral-related, analytic way of writing the stimulus properties of the brain event that precedes the operant response (R). It is the inferred cause of the operant response. After this *us* is paired with a reinforcer (S^{R+}/US), the response (*cr*) of excitement anticipates the excitement elicited by the (S^{R+}/US). It is a simple classical conditioning step that makes the *us* salient. The new compound is written *us-cr*, but in the brain, we do not differentiate the stimuli and responses. A stimulus for one layer is a response in another layer. The salience of the *us* is maintained by its continued pairing with the final reinforcer (S^{R+}/US). Thus, there is no extinction for responses to the *us* while these pairings are occurring. Over long periods of time a constant reinforcement rate loses its salience and responding becomes automatic.

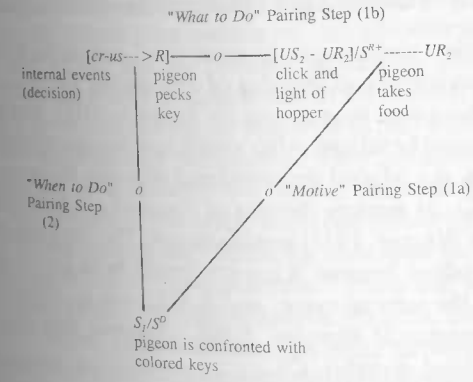
The value associated with an environmental stimulus outside (a red or a green light or a key) and the *cr* that represents the value associated with the density of past reinforcement are compared to the density of presently obtained reinforcement. If the values are discrepant, this elicits an emotional response that again is associated with each stimulus, the *us* and the S_1. This raises the salience of the *us* leading to changes in the vigour of one of the responses.

Partly from these observations, we make the following claim. Provided the *cr*

is salient, the animals are "aware" of what they are going to do before they do it. We speculate, and there is good evidence, that as one goes up phylogenetically, the planning step, or internalization of this unconditioned response, is what happens and becomes more regularized.

Once this pairing step occurs regularly, the environmental aspect elicits the response directly. This is why the response occurs more and becomes habituated. If there is no change in the value of the reinforcer, the subject no longer detects it unless there is a mismatch. For example, if you drive your car a particular route every day, you will not recall many of the details of the surroundings through which you drive. The drive is chunked into large segments, and the details control automatic driving but do not require reflective attention. If, however, you become lost, you suddenly become much more aware of your surroundings. This does not differ from having higher order verbal terms, such as long words, in which one does not necessarily plan or reflect upon each syllable. If the spelling is to be checked, the syllables are then examined.

An example will serve as an anchor for the theory to follow. A pigeon is presented with two keys in a Skinner Box for the first time. The pigeon sees grain in the lit hopper, and hears the click associated with its activation (an external stimulus, S_1). The sound and sight of the lit hopper (NS/S_1) are *paired* with the grain's consumption (US_2). The key peck is the response, and each key color becomes associated with one rate of reinforcement. When a change in reinforcement density takes place, the value of the key color changes. The conditioning of the sight of the grain (CS) comes to produce the conditioned response, that is, excitement (CR_2).



(3)

The "What to Do" Pairing Step (1b): Salience

Return to the example cited earlier just before the US becomes salient as diagrammed in expression three. The salience of the decision to peck the key (US) is established by pairing it with the arrival of the hopper and ingesting the grain (US_2/S^{R+}). The result of this pairing is that the decision to peck the key

(*US*) is now salient, because the pairing elicits the conditioned response and the stimulus it produces, (CR_2/CS_2) as diagrammed in expression four.

$$cr-(US-CR_2/CS_2)-R$$
excitement, (pecks the colored key, possibly more frequently)
image of the reinforcer

Before conditioning, these responses (CR_2/CS_2) were elicited only by the arrival of the hopper and the ingestion of food (US_2).

This proposed pairing step 1b has been anticipated by others. In Guthrie's (1952) theory, one role of the final reinforcing US_2 is to increase salience. In Tolman's (1932) theory, the *US* comes to elicit a portion of the responses to the reinforcing US_2. That is, the activity of preparing to act (*US*) elicits a set of memories (CR_2/CS_2) of past reinforcements and of the relationship (a relational event) between that previous activity and the reinforcements. For Tolman, this set of recalled events leads to the action being experienced as *purposeful* action. Our interpretation is slightly different. As subjects prepare to act (decide to peck the colored key), they recall the outcome of similar actions in the past (possibly get excited, salivate, and have an image of the food and its taste). These recalled events and the sense of the relationship between the previous similar actions and previous outcomes is experienced as the reason to act. Thus, in Tolman's terms, subjects have the sense that their acts are purposeful and voluntary, although this is actually an illusion, because memory is elicited, but is not in fact the cause of the act.

The "When to Do" Pairing Step: Environmental Control

In order for respondent conditioning to occur (i.e., so that pecking a colored key, *R*, follows the environmental stimulus of introducing the grain, (S_1), both the internal stimulus that elicits the reaching for the grain, (*US*), and the environmental stimulus, (S_1), must be salient. After a sufficient number of pairings during the "What to Do" step, the internal unconditioned stimulus becomes the salient complex, ($US\text{-}CR_2/CS_2$), as modern theories of classical conditioning (Mackintosh, 1974; Rescorla & Wagner, 1972) would suggest. The environmental stimulus, (S_1), will also be salient because it has previously been paired (pairing 1a) with (US_2/S^{R+}). Thus, the internal event, $us\text{-}CR_2/CS_2$, can be effectively paired with the neutral environmental stimulus, S_1/NS, changing that environmental stimulus into a conditioned stimulus, CS_1.

$$CS/S^D \text{ ----------} CR_2\ (R) \text{ ----------} US_2/S^{R+} \qquad (5)$$
perception of colored keys — arrival of hopper, pigeon gets food — frequent key pecks

The "When to Do" pairing results in a response following the environmental stimulus, S_1. Thus, the probability of pecking a given key is increased because

the complex that elicits the response is now elicited by the environmental stimulus. Changes in preference when reinforcement schedules are altered should depend on this step.

Results of the Three Pairings

The three pairings result in the following chain of events:

$$CS_1/S^D-(us-CR_2/CS_2)-R. \ldots\ldots US_2/S^{R+} \qquad (6)$$

After conditioning, the next time that the pigeon is confronted with the lit keys, CS_1, it will elicit the complex $us\text{-}CR_2/CS_2$, of repeated pecks to a given key. The model allows for only strengthening of responding. For example, this model predicts that punishment strengthens responding. The responses, however, are those that compete with the behavior that is being punished (Alkon, personal communication, January 23, 1990; Hull, 1952). Increasing the rate of making the competing response, increases the value of the outcome by decreasing the amount of punishment.

Extension to Two-key Preference

The reduction of the acquisition of a single operant response to respondent condition can be extended to the acquisition of preference. In the more complex two key situation, the role of value of the final reinforcer is emphasized rather than just its existence.

With this view of conditioning in mind, the Bailey and Mazur data were examined with the hope of developing a pilot model that could be simulated by a neural network. These pilot models are our first attempt to model preference acquisition and do not test the notions that they are based on. Commons and Hallinan (1990) use a very general notion of such neural nets. The interest here is to present a set of models that are restricted to the "pigeon view of preference" removing the experimenter perspective. They postulated that any frequently repeated response to a stimulus will lead to the slow development of a "cell assembly" within parts of the brain. Such cell assemblies are capable of acting briefly as a closed system and interacting with other such systems. A series of such events constitutes a "phase sequence." The phase sequence constitutes the melioration (Herrnstein, 1982; Herrnstein & Vaughan, 1980) or learning process.

In the two choice preference situation here, learning is postulated to occur when the amount of reinforcement delivered for responding on each key is altered. As with Grossberg's Adaptive Resonance Theory (ART) model (1980; chap 4 in this volume), the change in local reinforcement density creates a discrepancy between the value of reinforcement obtained historically over the short range and that obtained momentarily. When the reinforcement density changes, there is a discrepancy between the *CR*, the reinforcement predicted by the *us*, and the obtained density. The *us* elicits some motivational *CR*, which some consider an

expectation of a given reinforcement density. Tendency to respond to the little (*us*) also reflects an expectation of what the payoff for the response should be. The discrimination of this discrepancy makes the *us* salient again. This *us* is then paired again with the environmental S_1 that is more potent because it has been paired with the (US_2/S^{R+}). Together, the more-often-reinforced response is more vigorously elicited (Estes, 1969; Grossberg, 1982a, 1982b).

There is substantial evidence for the relativization of responding—that is the ratio of response allocation to one key to the response allocation of the sum of the two keys (e.g. Commons, Herrnstein, & Rachlin, 1982) as well as for individual strengthening of each response in the two key situation (Myerson & Mizzen, 1980). The sum of the rates increases with an increase in total reinforcement. The relative rates stay constant if the relative rate of reinforcement stays constant.

The Effect of the Time Between Reinforcement and a Response

Three time-related variables (Commons, Woodford & Trudeau, in press) affect acquisition. Each is identified with the effect that reinforcers have on a choice of behavior over time. The first is time scheduled associativity (Commons, Woodford, Boitano, Ducheny, & Peck, 1982), or the time lapse from each reinforcer to the choice that it affects. Another is the relative time between reinforcers, or the change in time lapse between reinforcers over several consecutive reinforcer/choice cycles. The assumption here is that the subject would choose more often the key with the shorter time delay between reinforcement. The last is based on the number of intervening events between reinforcers and choice. Time in this case is measured in terms of events passed rather than seconds.

Time allocation for responding to changes in scheduled reinforcement is a mechanism for characterizing melioration or learning. The value of events to a subject is reflected in how it spends its time (Commons, Woodford, & Ducheny, 1982). This model hopes to demonstrate that events in time and their value interact to determine allocation according to the matching law. The level of this allocation can be determined by response to scheduled reinforcement.

Our pilot models for acquisition are also based upon the General Additive Noise Model developed by Woodford (Commons, Woodford, & Ducheny, 1982). They describe the memory of an event, such as reinforcement, as some memory value plus some random noise term. For each subsequent event, an additional noise term is added to previous memories, so that over time, memories become more indistinct. For a four cycle experiment, for example, the model might predict the following memory values when the time came to make a choice. Here, M_i, is the memory of whether or not there was a reinforcer on cycle i, and n_i is the noise term associated with that cycle.

Commons, Woodford, and Trudeau (in press) have shown that this linear noise model predicts the hyperbolic decrease of the effect of reinforcers over time.

Cycle:	C_1	C_2	C_3	C_4 →	choice
Reinforcer:	1 (M_1)	1 (M_2)	1 (M_3)	0 (M_4)	
Memory 1:		M_1+n_1	$M_1+n_1+n_2$	$M_1+n_1+n_2+n_3$	$M_1+n_1+n_2+n_3+n_4$
Memory 2:			M_2+n_2	$M_2+n_2+n_3$	$M_2+n_2+n_3+n_4$
Memory 3:				M_3+n_3	$M_3+n_3+n_4$
Memory 4:					M_4+n_4 (7)

Further analysis has been carried out, examining the effect of the length of a cycle, measured in seconds. With that variable, the rate at which the remembered value of a reinforcement decreases, can be adjusted according to the average rate that events occur in a particular experimental situation. Hence, it is not how much time has passed that causes the decrease in value, but simply length of time relative to the average time between events in the particular environment. According to this hypothesis, a change in the length of the average reinforcement delay would have no effect, but a particularly long delay after the subject has already calibrated its rate of memory decay to a different rate of occurrence of events will affect the size of the reinforcer. In other words, the subject will give more weight to a reinforcer given after a relatively large time delay between reinforcers, remembering the most recent reinforcer for a longer period of time. In the interests of minimizing delay between reinforcement, the subject would develop a preference of the schedule that delivered reinforcement with the least amount of intervening relative time.

Using this model for decrementation of reinforcers over time, this simulation attempts to develop a learning algorithm that captures the time allocation predicted by melioration (Herrnstein, 1982; Herrnstein & Vaughan, 1980). Melioration predicts that the subject will modify its behavior so that the relative allocation of behavior matches the relative obtained reinforcement. Here, with two ratio-like schedules, subjects' behavior will stabilize on the richest reinforcement schedule over time.

Our study focuses on determining what happens when the density of reinforcement is changed and how the time allocation of choices that result can be predicted by the acquisition and delay functions (Commons, Mazur, Nevin, & Rachlin, 1987). The analysis of Bailey and Mazur's data presented here deviates from all past models including their own. Our pilot models examine reinforcer value over time instead of developing that fit based on the entire data set. This general model is used to calculate the total value of each reinforcer, each decremented for time, that has been delivered up to a given point in the experiment. The sum of the decremented values of all previous reinforcers yields the total reinforcement at any given point in time. This sum of all previous reinforcer values is then used to predict the behavior of the subject on the succeeding choice period. This process is repeated for each choice period, reassessing the value of reinforcers at each succeeding point in the experiment. We compared these predictions to data.

Method

Four pigeons were run on several 800 trial sub-experiments in a standard Skinner box. In each experiment, two keys were transilluminated, one with a red light and one with a green light. A peck on either key was reinforced with a single pellet of food, or not reinforced at all, with a probability based on which key was selected.

The probability of reinforcement on either key varied from trial to trial. Five different probability pairs were each run twice, with the rich or higher probability on one key first and then on the other. Probabilities were selected to test the effects of greater or lesser difference in probability, ratio between probabilities, and discrimination between values of probabilities. Before each trial, each bird was run on a special series of trials designed to bring the probability of pecking a given key as close to 0.5 as possible, and then a change in reinforcement density was slowly introduced. The data we are using are the 800 trial experimental sessions that were run after the transition sessions.

The Pilot Models

In the Linear Noise Model, the effect of any given reinforcer decreases over time as a decrementing hyperbolic function, as shown in figure 8.2. The curves represent the decrease of the two reinforcer values of R_1 and R_2, delivered at time t_i and t_{i+1}, respectively. The total value of both reinforcers at time T is the sum of the two decrementation functions at time T. The difference in time between time t_i and t_{i+1} is the relative time between reinforcers, denoted $deltat_i$. Then the total value at time t_{i+1} of R1 can be expressed:

$$R_1(t_{i+1}) = \frac{R_1(t_i)}{(t_{i+1} - t_i)}; \tag{8}$$

where $R_1(t_i)$ is the weight assigned to the reinforcer at the time it is obtained, and $(t_{i+1} - t_i)$ is the time elapsed at t_{i+1} since the reinforcer R1 was delivered. One possible value for $R_1(t_i)$ is the inverse of the time delay between R_i and the previous reinforcer, or $deltat_{i-1}$. This inverse would place a greater weight on reinforcers that were delivered with little intervening time. In this manner, reinforcers preceded by large time delays contribute less to memory, as one model for decrementation (Commons, Woodford, & Trudeau, in press) suggests.

Similarly, the total value of all reinforcers at time T can be expressed as the hyperbolically decremented sum of the values of all reinforcers delivered before T. That is:

$$\begin{matrix}\text{Total} \\ \text{reinforcement} \\ \text{at time T}\end{matrix} = \sum_{j=1}^{n} \frac{R_j(t_j)}{(T - t_j)}; \tag{9}$$

LINEAR NOISE MODEL

The Effects of Any Given Reinforcer Decreased Over Time Hyperbolically

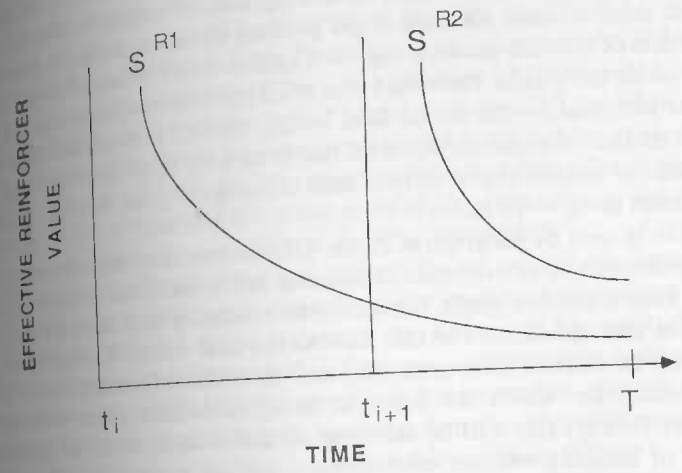


FIG. 8.2. Linear noise model. The effects of any given reinforcer decreased over time.

where n is the number of reinforcers that have been delivered up to time T, each $R_j(t_j)$ is the value of the jth reinforcer when it was delivered at time t_j, and the inverse of $(T - t_j)$ is the hyperbolic decrementation of the jth reinforcer since it was delivered at time t_j.

Because the value of each reinforcer decrements hyperbolically, the contribution of any reinforcer at the moment it is delivered is infinite. This reflects the certainty of how the subject will choose at the moment the choice is made. Immediately thereafter, the contribution begins to drop off. Because the contribution is infinite at the moment the reinforcer is delivered, its value cannot be calculated. Thus, the difference of the values of the reinforcer at the moment of delivery to its contribution to behavior one time unit later, is similar to Mazur's (1987) adding a constant.

The simplest model assumes that the value of any reinforcer when it is given $(R_1(t_i))$ is 1. Then 1 time unit after the reinforcer is given the value of the reinforcer can be written:

$$R_1(t_{i+1}) = \frac{R_1(t_i)}{(t_{i+1} - t_i)} = \frac{1}{(x + 1 - x)} = 1; \tag{10}$$

where x represents the time that the reinforcer was delivered. Two time units away, this value would be 1/2, three time units away 1/3, and so on. The total value of several such reinforcers would be the sum of the values of each of these reinforcers at some point in time.

This method of evaluating reinforcers is the unweighted sum of decrementation, so called because the value is unweighted, or simply 1. Graphically, a possible reinforcement schedule might produce the sums shown in Figure 8.3. Here, units of time are shown along the x axis, whereas values of reinforcers are shown along the y axis. The total value of all previous reinforcement at a given point is indicated by the dotted line, which, until t_{i+1}, is coincident with the decrementation of reinforcer R_i (since that is the only reinforcer operating). The total value of reinforcement 10 time units along is the point at which the dotted line crosses the x = 10 line.

As can be seen by the graph in Figure 8.3, the sum of all reinforcement yields a noncontinuous curve over units of time that can be described graphically as the sum of each reinforcer graph, but cannot be accurately evaluated for any given instant of reinforcement. For this reason, the total value of reinforcement is calculated in discrete time intervals, and the value of reinforcement at the time period, for which the value is being calculated never includes the reinforcer (if any) that will be delivered on that instant, since the subject has no way of knowing whether reinforcement will occur until immediately after

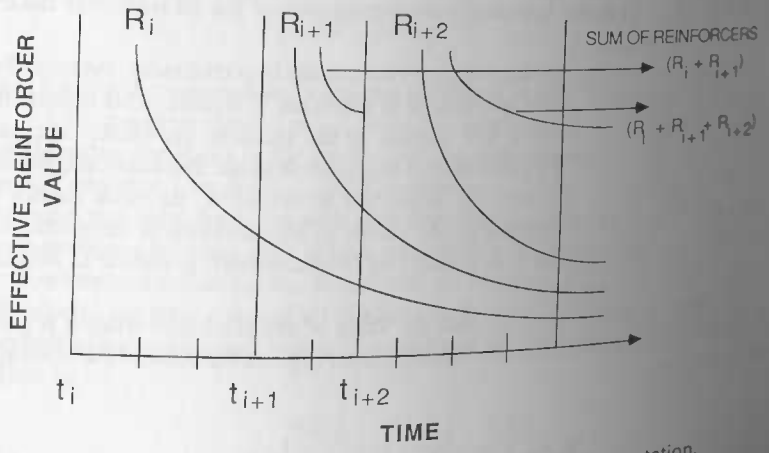


FIG. 8.3. Sample schedule for unweighted sum of decrementation.

the choice is made. This yields a smoother, discrete step function with an overall tendency of responding to increase as long as reinforcers are delivered with some regularity, but a local tendency to decrease as reinforcers decrement over discrete time intervals.

Graphically, a possible total reinforcement function is shown later for unweighted reinforcers. The bumps indicate that a reinforcer was delivered just prior to it, and give an overall increase to the function.

Note that if reinforcement were discontinued, the function would drop off, approaching but never reaching 0. Imagine a situation where the subject ceased to receive reinforcement for some type of behavior. Eventually, the subject would stop that type of behavior in favor of some type that did deliver reinforcement with some regularity, if possible. The subject would, however, always remember that the first type of behavior did at one point elicit reinforcement, and thus the memory of reinforcement would never theoretically reach 0.

The discrete time interval that was chosen for the task of calculating these curves in the pilot models we evaluated was the number of intervening events. For this experiment, this is equivalent to the number of key pecks between reinforcers. The use of this time scale solves two problems. First is the question of which time scale most accurately represents decrementation, and second is the role played by the number of intervening events from the original decrementation model. Both of these questions will be more fully addressed later.

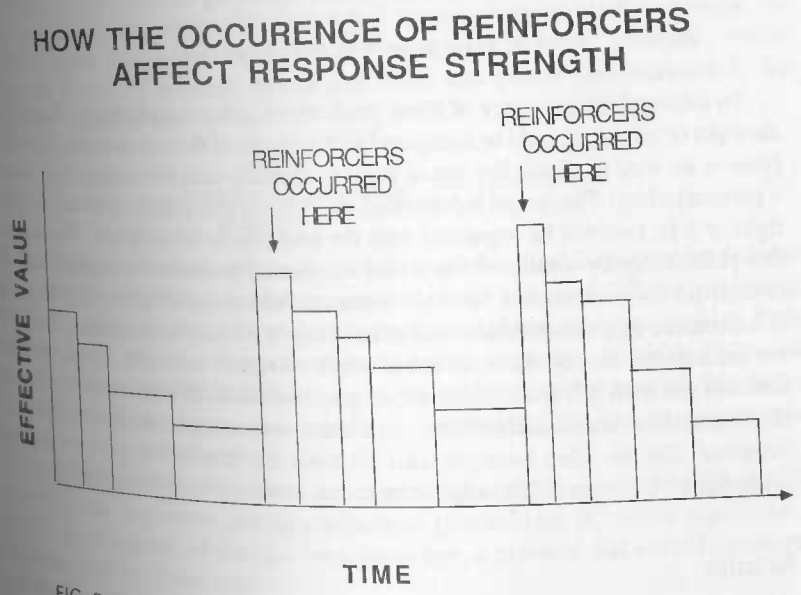


FIG. 8.4. How the occurrence of reinforcers affect response strength.

'latency' (the delay between lighting the right and left keys and the bird's choice of one) was ignored, and only the trial itself was counted as a time unit. In this manner, the weight of each reinforcer was given a larger overall value. This occurred because there were, as a rule, more tenths of a second between reinforcers than there were events, and the reciprocal of these values for tenths of a second was smaller:

Weighting term for model:

$$\#1\ \frac{1}{\text{tenths of a second}} < \#2\ \frac{1}{\text{events}} \tag{15}$$

These regressions on the whole varied little from those based on the first model because the weighting terms differed but not significantly.

More interesting are the pilot models run without the weighting terms. These simply count each reinforcer as a '1' one time unit after they are delivered, and sum the hyperbolic decrementation of the reinforcers as described earlier. The most significant of these used events as time units, as with the second weighted model, cited earlier. For nearly all schedules, this model yielded better fits and more significant regressions R^2 values for this model range between 8% and 52%, with an average R^2 of 36%. A comparison is shown between Bailey's fits for predicting response to the rich key over time and our fits for predicting probability to peck right over local probability for the same schedules.

Due to the rapid decrementation of each reinforcer in the hyperbolic model, there is a heavy bias towards the most recent reinforcers in the estimation of the probability to peck right. For example, based on this model, a large string of right reinforcers may be nullified by a single succeeding left reinforcer, especially if there is a significant delay between these reinforcers. Subjects have a much greater tendency to discount these sparse left reinforcers, even though the model suggests that each should have as much weight as a right reinforcer. For this reason, the pilot model tends to 'overreact' to the experimental conditions, with two noticeable affects; the model tends to fluctuate over a wider band than the subject actually does, and the probabilities tend towards the extremes. This can be seen in scatterplot 1 (see Figure 8.5), where the predicted values are along the x-axis and the local probabilities are along the y-axis. Note how the data ranges from 0 to .45 with fairly normal distribution, but predictions for this band have values appearing primarily between 0 and 0.15. In this way, a variation of probabilities on the part of the subject is overemphasized by the model because the probabilities all lie below 0.5. The model gives this marginal preference undue weight, and reduces the prediction still lower to hover around 0.1.

The counterpart of this overemphasis by the model can be seen in the same graph by examining the range of the values on both axes. Whereas values of the local probabilities range from 0 to 0.45, predicted values for these probabilities, though centering at 0.1, range from 0 to 0.8.

That the model seems to emphasize the most recent reinforcers may be a

COMPARISON OF BAILEY'S ANALYSIS WITH THE PRESENT ANALYSIS

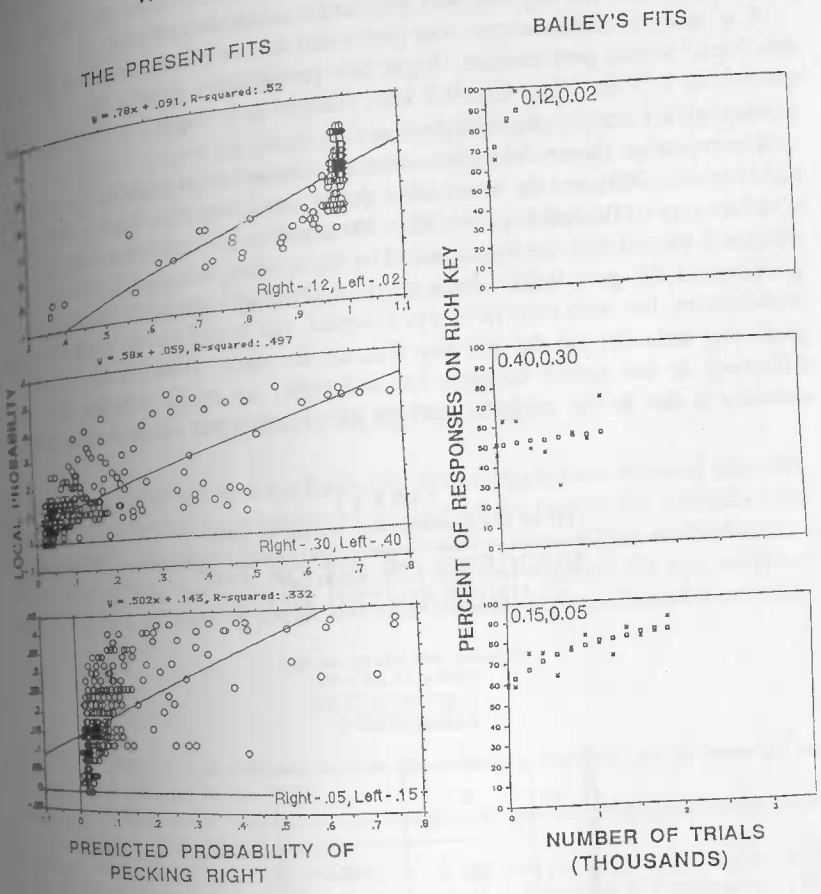


FIG. 8.5. Comparison of Bailey's analysis with present analysis.

function of the way the local probabilities are calculated. Because estimation of the local probability involves averaging over the behavior of the last fifty trials, local fluctuations in the subject's value assignments caused by recent reinforcement may not be given enough weight.

The fit was further complicated by two forms of bias—a global preference for a given key before reinforcement begins, and local patterns of behavior induced by the correction procedure. The first of these biases occurs when a tendency of the subject to prefer one key over the other is not totally eradicated by the correction procedure. The second occurs because the correction procedure nor-

malizes behavior on a global, but not a local scale. Although the correction procedure effectively moves the subject back to evenly distributed response on both keys, the method of alternate reinforcements creates a local tendency for the pigeon to choose the key that was not reinforced in the last trial.

A χ^2 test for independence was performed on our pilot model's predictions and birds' actual performance. Right key probabilities greater than 0.5 were counted as 1. Values less than 0.5 were counted as 0. Right key response were counted as a 1 and left key response as 0.

The top table shows data where the rich reinforcement schedule was on the right key (N=269) and the lower table shows data where the rich reinforcement schedule was on the left key (N=309). No evidence for dependence was found. $\chi^2(1) = 0.84$ and was not significant. The lower table, was significant however. $\chi^2(1) = 11.53$, $p < 0.001$. Because the tables show the same reinforcement probabilities, but with their rich keys reversed, our model seems to be better at predicting behavior on the left key than on the right. With no mathematical difference in our model between left and right, we might suppose that this anomaly is due to the subjects entering the experimental sessions with a bias

TABLE 8.1
Fit of the unweighted model to data

SUCCESS OF FIT OF THE UNWEIGHTED MODEL

Schedule with rich key on right
(right = .15, left = .05)

Predicted Choice	Actual Choice: 1	Actual Choice: 0
≥.5	6	13
<.5	56	194

Schedule with rich key on left
(right = .05, left = .15)

Predicted Choice	Actual Choice: 1	Actual Choice: 0
≥.5	249	37
<.5	14	9

toward the left key. In the upper table, where the rich schedule was on the right key, the majority of successful predictions were correct rejections. On the lower table, which displays predictions of behavior when the left key is rich, the majority of successful predictions were hits.

Conclusions

This investigation addressed questions on acquisition of preference. A proposed reduction of operant to respondent conditioning was extend to the two key preference situation. The Linear Noise Model was then combined with it to produce a pilot models explored here. The unique characteristics of these models included trial by trial analysis of the probability to peck a given key, analysis from the subject's view point and prediction of the future probability of pecking a given key.

ACKNOWLEDGMENTS

The data was collected by John Bailey for his undergraduate thesis at Harvard. He and James Mazur have submitted a report of the data to the Journal of the Experimental Analysis of Behavior. A condensed version of their method section is reproduced here with their permission. They also supplied the data analyzed here. We thank William Reynolds and Jared Jenisch for their editorial comments.

REFERENCES

Bailey, J. T. (1988). *Factors Affecting Pigeons' Development of Preference for the Better of Two Alternatives*. Harvard Honors Thesis.

Bailey, J. T., & Mazur, J. E. (submitted). Choice behavior in transition: Development of preference for higher probability of reinforcement.

Commons, M. L., & Hallinan, P. W. with Fong, W., & McCarthy, K. (in press). Intelligent pattern recognition: Hierarchical organization of concepts. In M. L. Commons, R. J. Herrnstein, S. M. Kosslyn, & D. B. Mumford (Eds.), *Models of behavior, 9, Computational and clinical approaches to pattern recognition and concept formation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Commons, M. L., Herrnstein, R. J., & Rachlin, H. (Eds.). (1982). *Quantitative analyses of behavior: Vol. 2. Matching and maximizing accounts*. Cambridge, MA: Ballinger.

Commons, M. L., Mazur, J. E., Nevin, J. A., & Rachlin, H. (1987). *Quantitative analyses of behavior: Vol. 5, Effect of delay and intervening events on value*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Commons, M. L., Woodford, M., Boitano, G. A., Ducheny, J. R., & Peck, J. R. (1982). Acquisition of preference during shifts between terminal links in concurrent chain schedules. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analyses of behavior: Vol. 3, Acquisition* (pp. 391–426). Cambridge, MA: Ballinger.

Commons, M. L., Woodford, M., & Ducheny, J. R. (1982). How reinforcers are aggregated in reinforcement-density discrimination and preference experiments. In M. L. Commons, R. J.