

SITUATIONAL DESCRIPTIONS OF BEHAVIORAL PROCEDURES: THE IN SITU TESTBED

STEVEN M. KEMP AND DAVID A. ECKERMAN

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL



We demonstrate the In Situ testbed, a system that aids in evaluating computational models of learning, including artificial neural networks. The testbed models contingencies of reinforcement using an extension of Mechner's (1959) notational system for the description of behavioral procedures. These contingencies are input to the model under test. The model's output is displayed as cumulative records. The cumulative record can then be compared to one produced by a pigeon exposed to the same contingencies. The testbed is tried with three published models of learning. Each model is exposed to up to three reinforcement schedules (testing ends when the model does not produce acceptable cumulative records): continuous reinforcement and extinction, fixed ratio, and fixed interval. The In Situ testbed appears to be a reliable and valid testing procedure for comparing models of learning.

Key words: neural networks, reinforcement schedules, situated action, learning theory, key peck, computer simulation, POMDP

The purpose of the present paper is to introduce a new technique, called an In Situ simulation, for evaluating and testing computational models of learning, including the recently developed neural networks (e.g., see Donahoe, Palmer, & Burgos, 1997). The testing is carried out by (a) exposing the learning model to schedules of reinforcement and (b) generating cumulative records of the learning model's output. These records may be compared to those a pigeon generates following comparable exposure. We hope that In Situ simulations will provide a standardized approach to testing learning models that seek to characterize operant learning to allow comparison across models. We consider that the In Situ testbed has a function much like a wind tunnel used in testing models of airplane design. The testbed provides the wind patterns, and the model shows how it flies in these wind patterns. In our case we are seeking to know if the model "flies" in a manner that is similar to a pigeon.¹

We thank Adam N. Rosenberg for his invaluable assistance with both mathematical and conceptual issues and J. E. R. Staddon, Ido Erev, and Gene Steinhauer for their assistance with the various models tested. We also thank the four anonymous reviewers for their guidance.

For additional information, reprints, and copies of materials, please contact Steven M. Kemp, Psychology Department, Davie Hall, CB 3270, University of North Carolina, Chapel Hill, North Carolina 27599-3270 (E-mail: steve_kemp@unc.edu).

¹We are currently implementing a version of the In Situ testbed for use in evaluating state-of-the-art models of learning. To make use of this resource, readers are

Quantitative approaches in psychology might be divided into two types. In one approach, quantitative methods are used to produce a general expression of a functional relation between behavior and the environment. The matching law (Baum, Schwendiman, & Bell, 1999; Herrnstein, 1997) is such a formulation. Each functional relation captures some aspect of a behavioral phenomenon expressed at its own level of analysis and is usually tied to a class of experimental procedures that demonstrate that phenomenon. Using the second approach, a quantitative system is constructed that accepts input from an environment and produces output that may be functionally related to this input. The procedure that generates the output is intended to capture the effects of the environment and the neural processes that produce real behavior. The level of analysis must correspond to the level of detail of the description of the behavioral procedure. The more effects captured, the more successfully control by the procedure is predicted. When the system

invited to send to the first author an executable program module for DOS or Windows (Windows NT compatible) that accepts input and provides output as described below. Cumulative records will be returned to the sender to begin an evaluation of his or her model. These cumulative records may be compared to records produced by a pigeon in procedures similar to our tests. Although the current implementation requires collaboration with the authors, a future stand-alone implementation is planned. This stand-alone implementation will be sufficiently user-friendly (and documented) for readers to evaluate their own models.

changes its output on the basis of a prior history of inputs, this may be called a *learning model*.

Skinner (1950) raised cautions regarding the use of learning models in his well-known essay titled “Are Theories of Learning Necessary?” He worried, for example, that interest in developing models might distract researchers from carrying out a careful experimental analysis of the behavior of real organisms. Although we agree with Skinner’s priority of “experimental analysis first,” we propose that the time is here to use a modeling approach to bridge the gap between the extensive experimental analysis that has accumulated under Skinner’s banner and the growing database of new neurophysiological data. The availability of these two databases makes possible the kind of model testing we espouse in this report.

Specifically, Skinner (1984a, 1988a) said the following about mathematical models:

No matter how many of the formulations derived from a study of a model eventually prove useful in describing reality . . . , the questions to which answers are most urgently needed concern the *correspondence between the two realms* [italics added]. How can we be sure that a model is a model of behavior? *What is behavior, and how is it to be analyzed and measured? What are the relevant features of the environment, and how are they to be measured and controlled?* [italics added] How are these two sets of variables related? The answers to these questions cannot be found by constructing models. (p. 514, p. 83)

We propose that the In Situ approach advocated in this report speaks directly to Skinner’s concerns. It offers a formal and systematic method for relating models to behavior by establishing a standard set of logical relations between the environmental and behavioral variables. To the degree that our situational view of how relevant features of the environment exert their control over behavior is correct and general, the models tested here will be compelled to demonstrate that they are models of behavior by coming under control of the simulated environmental contingencies in the same way that real behavior comes under the control of real contingencies. Further, with the relations between environmental and behavioral variables fixed, each model tested is challenged to demon-

strate that it is competitive with other models by coming under the control of the same environmental variables simulated in the same way. We offer this report realizing that our specific implementation of the relation between environment and behavior is incomplete. We believe, however, that we demonstrate what can be gained if the assumptions of *situativity theory* (Barwise & Perry, 1983; Clancey, 1993) are applied to describing how environmental and behavioral variables are related.

When Skinner (1950) presented his concern with models of learning, the coffers of neuroscience were nearly empty and theorists were free to fill their “conceptual nervous system” with anything needed by their theory. In the decades since 1950, however, neurobiological knowledge has increased considerably. We join Donahoe and Palmer in advocating that we include this knowledge into our thinking. We also join these authors in encouraging the use of “formal interpretation” (Donahoe & Palmer, 1994, pp. 128–129; Shull, 1995, pp. 350–353). Skinner (1984b, 1988b) provided a specific recommendation regarding the needs of behavior analysis for the data of neuroscience:

A behavioral analysis has two necessary but unfortunate gaps—the spatial gap between behavior and the variables of which it is a function and the temporal gap between the actions performed upon an organism and the often deferred changes in its behavior. These gaps can be filled only by neuroscience, and the sooner they are filled, the better. (p. 722, p. 470)

We propose that the In Situ approach advocated in this report will help in filling these gaps.

Although complex computational models of learning, especially neural networks, are growing in popularity (see references for the models we test below), there are many issues that make their evaluation difficult. One is that the specific predictions made by a model for various conditioning procedures are difficult to determine even when the equations are published. Whereas the predictions of older quantitative models often could be calculated directly with paper, calculator, and pencil, these newer, dynamic models produce predictions that are highly dependent on the particular sequence of events involved. Our

In Situ approach provides a means of generating these predictions in a standard, repeatable, and documentable manner, more or less independent of the model chosen.

Another difficulty is that published predictions for learning models are often expressed in terms of behavioral measures that are convenient for the modeler and specific to the problem or phenomenon the modeler is addressing. These predictions may be difficult for the reader to evaluate as part of a general comparison of models. In fact, because each model is presented with different procedures and different measures, the comparison of competing models is especially difficult. Our In Situ approach allows a wide variety of learning models to be swapped in and out and uses a standard set of testing procedures with each, thus facilitating model comparison.

In developing our In Situ testbed, we are following our own advice (Kemp & Eckerman, 1995) as well as that of Church (1997). We are aware of no published simulations of any computational model of learning that implement more than a small number of the features recommended by these authors for making direct comparisons of different learning models. We offer such comparisons below.

The In Situ approach conforms to all of Kemp and Eckerman's (1995) and Church's (1997) recommendations listed here:

1. Separate models from the behavioral procedures. We recommend building learning models and behavioral procedures into separate software modules and setting standards for data passed from the test procedure to the model (the stimuli) and data passed from model to procedure (the responses). This allows a set of multiple procedures to be used to test any model and multiple models to be tested with each procedure. Testing can be accomplished simply by swapping modules in the simulator.

2. Record model outputs in a form that is comparable to records of real behavior. If output record formats are standardized to match those from experiments using real organisms, this allows an improved comparison between the behavior of the synthetic and the real organisms. Church (1997) specifically recommends a "Time.Event" format recording "the sequence of times of occurrence of

stimuli and responses" (p. 387). This approach also allows a variety of different behavioral measures to be summarized identically from both the real organism and the model (e.g., ratios of responses, etc.). This is the approach Kemp and Eckerman (1995) called a "direct analysis."

3. Use a standard set of multiple graded procedures to evaluate each model. Church (1997) advocates the use of multiple behavioral procedures to evaluate any single model. Here, we go further by recommending that each model be subjected to a standard sequence of procedures, graded from easy to hard. In the present study, we use a sequence of single-key operant procedures: continuous reinforcement (CRF), continuous reinforcement followed by extinction (CRF/EXT), fixed-ratio (FR) reinforcement, and fixed-interval (FI) reinforcement. Using a standard sequence facilitates crossmodel comparison. Using a graded sequence helps to pinpoint functional shortcomings of the model. If a model passes an easier test and fails a harder one, that suggests that challenges added in the harder test required behavioral capacities beyond the scope of the model. Specific reasons for choosing different reinforcement schedules to comprise this series are noted in the discussion at the end of this report.

4. Situate responses and stimuli spatially as well as temporally. Kemp and Eckerman (1995) advocated that descriptions of learning models and behavioral procedures represent the organism and its environment as located in both space and time. Recently, situativity theory (Clancey, 1993, 1997; D. A. Norman, 1993; Suchman, 1993; Vera & Simon, 1993a) has emphasized the many important implications that follow from this proposal. With respect to stimuli, for example, what a pigeon sees (i.e., is visually controlled by) depends critically on where it is looking. The display available to an organism at a particular place and time is called a *situated stimulus*. The timers and counters of the reinforcement contingency are not available to the pigeon and are not included in the situated stimulus. Below we will connect this characteristic with an approach to describing environmental contingencies named *partially observable Markov decision processes* (POMDPs).

Analysis in terms of operants, describing responses in terms of their effects, demands

that movements made by an organism (e.g., a head thrust) also must be situated. The closure of the key switch, for example, can only be accurately simulated by taking into account both the behavior and the location of the organism at the time of emission.

The issue of situativity is relevant to the gaps noted by Skinner. A neuroscientific theory cannot simply be plugged into an environment and treated as a model of behavior. Although a neural network or other biobehavioral model may make a simulated body move (just as real neural activity makes a real body move), these bodily movements in turn must be situated to bring about changes in the environment (e.g., a bar press, a key peck). Typically, behavioral accounts are given in terms of environmental circumstances and behavioral accomplishments (functional terms). Neuroscientific accounts, however, are given in terms of classes of neural events and bodily movements (formal or topographic terms). A translation must be made between bodily movements and the accomplishments they bring about (i.e., operants). This dilemma, which we call the *neural network dilemma*, was examined in detail by Weiss (1924, pp. 42–44, 1925, pp. 55–56; see also Guthrie, 1940, and Lee, 1983, 1986, 1996). The In Situ approach provides a means of solving this dilemma, thereby permitting the integration of behavioral and neuroscientific data.

5. When comparing models, first compare their output for each procedure to that of a real organism and then compare their relative successes (in matching real behavior) to determine the best fit. Testing learning models using the In Situ testbed facilitates model comparison, but it does make things tougher on the individual models. Traditionally, authors of computational models have used tests that show their models to best advantage by selecting and designing procedures appropriately. The model is often developed to make concrete a general point being raised by the author. To emphasize this point, the model is tested in a restricted and specific way.² The successful initial simulation merits publication and introduces the model to the

scientific community. If the new model is to have a broad impact on the field, however, it must be tested in a manner that permits comparisons with other models. Our use of fixed standardized sets of procedures makes this possible. Because our tests are more demanding, the models tested here will fare rather less well than in their previous publications.

The present study is designed to demonstrate the reliability and validity of the In Situ testbed. To accomplish this goal, three published models of learning (Daly & Daly, 1982; Roth & Erev, 1995; Staddon & Zhang, 1991) were compared under four reinforcement conditions (tasks): operant level, CRF followed by EXT (both in Simulation 1), FR reinforcement (in Simulation 2), and FI reinforcement (in Simulation 3). These tasks were chosen independently of the strengths or weaknesses of any particular model. Instead, the tasks were chosen as those producing (in real animals) the kind of results a broadly applicable learning model should aspire to predict. In a sense, the tasks are standardized tests for learning theories.

Despite these restrictions, model parameters were freely adjusted to produce the best possible performance by each learning model. Because the testing procedures were standardized, the adjustments to each model to optimize results did not produce unfair advantage over other models. We acknowledge that establishing standards is a risky business because it can close off divergent paths. We invite the reader to consider the In Situ testbed as a starting place for the development of appropriate standards.

Mechner's Diagrams As a Base for Our Work

A brief tutorial on Mechner's system for describing reinforcement schedules is provided, because this system is the starting point for the diagrams we use for specifying the simulated environmental contingencies within the In Situ testbed. Mechner diagrams (Mechner, 1959; Millenson, 1967; Weingarten & Mechner, 1966) facilitate model comparison by standardizing the precise specification of procedures in a publicly accessible format. Figure 1 shows Mechner diagrams for CRF, FR reinforcement, and FI reinforcement schedules. Each square bracket indicates a distinct state in the contingency. The stimulus symbol

² Noteworthy exceptions include Daly and Daly (1982) and Schmajuk, Lamoureaux, and Holland (1998), in which an attempt is made to test the model using a comprehensive sample of behavioral procedures, albeit using nonsituational representations of those procedures.

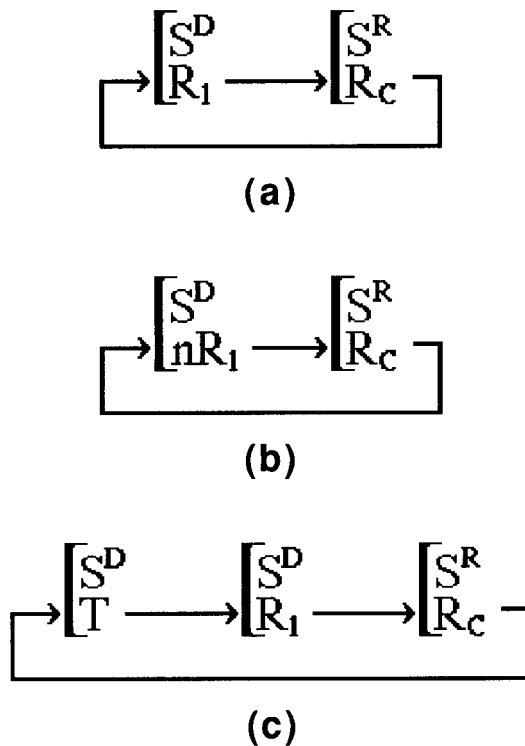


Fig. 1. Mechner diagrams for (a) CRF, (b) FR reinforcement, and (c) FI reinforcement schedules. Each square bracket indicates a distinct situation. The stimulus symbol (S) at the top of each bracket indicates the entire prevailing stimulus complex for that situation. S^D indicates the presentation of the discriminative stimulus. S^R indicates the presentation of the reinforcing stimulus. Each arrow indicates a change in situation. The symbol at the tail of each arrow (R for responses, T for time elapsed) indicates the event that precipitates the change. R_1 indicates the operant response. R_c indicates the consummatory response. n indicates that the following response must be emitted multiple times. T indicates that a period of time must pass.

at the top of the bracket indicates the prevailing stimulus complex for that state. The term S^D indicates presentation of the discriminative stimulus; S^R indicates presentation of the reinforcing stimulus. Each arrow indicates a change in state: The symbol at the tail of the arrow indicates the event that precipitated the change. R_1 indicates the operant response; R_c indicates the consummatory response. Thus for CRF (a), a single operant response in the presence of S^D produces the reinforcing stimulus and allows the consummatory response to return the procedure to the initial state. FR reinforcement (b) is similar except that the term n indicates that the

operant response must be emitted n times to shift state. When FI reinforcement (c) is arranged, T seconds must elapse in the presence of the S^D before the state changes to one in which an operant response produces the reinforcing stimulus.

The In Situ testbed expands Mechner's (1959) system as we describe below. We also integrate principles of situational semantics (Barwise, 1989; Barwise & Perry, 1983). This extension establishes a connection between reinforcement schedules and a formal class of mathematical entities called POMDPs³ (Jaakola, Singh, & Jordan, 1995; Kaelbling, Littman, & Cassandra, 1998; Monahan, 1982; Singh, Jaakola, & Jordan, 1994; Sondik, 1971). Use of POMDPs further facilitates crosslaboratory replicability as well as the implementation of certain situational features (see below). Also, the use of POMDPs places In Situ simulations firmly within the framework of the computational learning approach developed by Sutton and Barto (1998) that is known in engineering circles as *reinforcement learning*.

SIMULATION 1: CONTINUOUS REINFORCEMENT

METHOD

Models

To demonstrate the capabilities of the In Situ testbed, three models were selected to serve as subjects of the simulations. The criteria for inclusion were that (a) the model is claimed to account for operant behavior, (b) claims for the model are confirmed by the results of published simulations, and (c) the model was published in enough detail to allow reprogramming into the testbed. In addition, the models differ substantially from one another in the mathematical techniques employed in the calculations.

Each of the three models is summarized below. Each original model had to be augmented to implement it within the In Situ testbed. For purposes of clarity, we will designate each original unnamed model using

³ The well-known limitations of Markov models as models of learning (Kintsch, 1970) are not of special concern here. Our claim is that POMDPs are very general representations of behavioral procedures, *not* of learning.

the last names of its authors and designate our implementation using their initials. When the original model was named by its authors, we will add an asterisk to designate our implementation. Although presentation of the details for these models is not necessary to our present purpose, we have included a full description of our implementations in Appendixes A, B, and C for purposes of documentation and replication.

SZ model. Staddon and Zhang (1991) offer a simple stochastic model of reinforcement effects, selecting the behavior (of m potential behaviors) with the highest calculated value (V) on each cycle, the winner-take-all method. We call our implementation of that model the SZ model. The model has two parameters, a and b . The a parameter (adaptation or conditionability) determines the length of time the effect of reinforcement persists. The b parameter (arousal or elicibility) determines the magnitude of the effect of reinforcement. Each behavior's V value on each cycle is a stochastic function of a and b and $V(t - 1)$, its value on the previous cycle. Stochasticity insures that the maximum V varies from step to step.

The equations for the SZ model implemented in the present tests specified two details that were left indefinite in the published article. The original theory included no explicit model of antecedent stimuli. In our equations, a separate V value was created for every possible stimulus–response pair. Further, initial V values were set to .5 instead of 1, because this provided better initial stability. Appendix A provides the detailed equations used in the simulations of the SZ model.

RE model. Roth and Erev (1995) offer a simple account of selection by reinforcement that also calculates a V value (always nonnegative) for each behavior on each iteration. Here, the likelihood of a behavior being emitted on a given iteration is proportional to the ratio of its V value to the sum of all the V values (Luce, 1959). The V value of the behavior emitted just prior to reinforcement is augmented by 1. Roth and Erev tested their algorithm in trial-by-trial simulations, but no specific guidance was provided for generating a moment-to-moment account. Our implementation, called the RE model, thus adds an exponential decay function to the original model to account for the effects of delay of

reinforcement. As with the SZ model, a V value was created for every possible stimulus–response pair rather than just for each response. The reinforcer altered all V values, but the likelihood of emitting a response was calculated only with respect to the V values for the immediately present stimulus.

There are two parameters: *reinforcer potency* and *decay rate* of the effects of the reinforcer over time since the response. The net effect of reinforcer potency was halved after each second, following Staddon and Zhang (1991). Reinforcer potency was set to 1.0 (arbitrarily, because only one type of reinforcer is utilized). Appendix B provides the detailed equations used in the simulations of the RE model.

DMOD model.* Daly and Daly (1982, 1984) modified Rescorla and Wagner's (1972) model of classical conditioning, such that it could be used to model instrumental conditioning. They labeled their model DMOD. We label our implementation the DMOD* model.

The original model assigns a V value for each stimulus–response pair, which is recalculated at each iteration. Each V value is the sum of three components: *approach*, *avoidance*, and *counterconditioning*. (a) The approach component is calculated according to the Rescorla–Wagner (1972) theory. The approach value varies between two asymptotes, 0 and λ (Bush & Mosteller, 1951, 1955; Narendra & Thathachar, 1989). The value increases when the reinforcer is delivered and decreases when the reinforcer is not presented. (b) The avoidance component is a negative number that ranges from 0 down to twice the arithmetic inverse of the approach value. It is included to represent a diminished influence of the reinforcer when a difference exists between expected and actual reinforcer amount. (c) The counterconditioning component moderates the avoidance value. It is positive and ranges between 0 and the arithmetic inverse of the avoidance component, rising as reinforcer amount increases and falling as reinforcer amount decreases.

As with the Rescorla–Wagner model, DMOD and DMOD* contain two types of parameters, α and β . α indicates the salience of each of the discriminative cues. β indicates distinct learning rates for the different components of V . Each of the three components—approach, avoidance, and counter-

conditioning—has two β s, one for increasing V and one for decreasing V , for a total of six β s in all. In addition, there are β s for secondary reinforcement that can be set separately from those for primary reinforcement, making for 12 β parameters in all. (Usually, most of the β parameters are set equal to one another, making for only two to four parameters.)

There are important differences between DMOD and the Rescorla–Wagner model as well as between DMOD and DMOD*. The Rescorla–Wagner model can be interpreted as predicting changes in associative strength between stimuli due to the presentation of what has been labeled by some as a reinforcer, the unconditioned stimulus (Mackintosh, 1983, pp. 189–192). On this reading, the occurrence of the conditioned response is an index of the degree of associative strength of the conditioned stimulus (Mackintosh, pp. 19–22). In like fashion, Daly and Daly (1982) used instrumental responding as an index of associative strength of discriminative stimuli due to the subsequent presentation of a reinforcer. In addition, secondary reinforcement is modeled by having each discriminative stimulus acquire reinforcing properties proportionate to the V values conditioned to it (Sutton & Barto, 1998).

This approach leaves open two questions regarding concurrent responses. Given the associative strengths at a particular moment in time, what determines which of the available responses is emitted? Given that a particular response is emitted, which associative strength should be altered due to reinforcement? Daly and Daly (1982, pp. 465–466) explicitly avoid the first issue in their discussion of choice, but their discussion indicates how they handle the second issue. In essence, associative strengths for choice alternatives are calculated separately and are then compared. In their trial-by-trial simulations, Daly and Daly reported the differences between associative strengths as the dependent variable. In implementing DMOD in moment-to-moment simulations, Steinhauer (1986, chap. 7) follows this approach by constructing a separate system of associative strengths for each alternative response. After a response is emitted, V values associating that response (and only that response) with the preceding stimuli are altered due to any subsequent primary and

secondary reinforcers. As with the Rescorla–Wagner model, all occurrent stimuli share an asymptote. Any changes in associative strength, however, apply only to the response emitted.

Our DMOD* model follows this same approach. Whereas for the SZ and RE models, a separate instantiation of the model was added for each stimulus, for DMOD* a separate instantiation was added for each response. DMOD* resolves the first issue in a manner similar to the RE model. V values for all occurring stimuli are summed for each response. The probability of each response is the total of the V values for that response, divided by the sum of all the positive totals (Luce, 1959). Negative totals are ignored.

Finally, the addition of multiple responses to DMOD creates ambiguity in how to handle secondary reinforcement. In DMOD, the reinforcing efficacy of a stimulus cue is determined relative to the V value for that cue. In DMOD*, each cue has a separate V value for each response. These V values must be combined to calculate the overall reinforcing efficacy of the cue. DMOD* uses the V values for each response to construct normalized weights, which are then used to determine each of the asymptotes for secondary reinforcement of the preceding cues. Appendix C provides the detailed equations used in the simulations of the DMOD* model.

Apparatus

Hardware and software. The current version of the In Situ testbed is programmed in SAS/IML® (SAS Institute, 1996), and simulations are performed on a Dell 450DE, a standard Intel® 486-based desktop personal computer, with the OS/2® Warp operating system.

Testbed design. Overall, the In Situ testbed consists of three parts: (a) the testbed proper, which simulates the functioning of the operant chamber and a minimal set of behavioral topographies of the pigeon; (b) the learning model under test, which interacts with the testbed proper to generate behavioral protocols; and (c) the graphics routine, which converts the behavioral protocol into a cumulative record.

The interaction between testbed and model derives from a general approach called *dynamic programming* (Bellman, 1957; J. M. Norman, 1975). The testbed constitutes a Markov

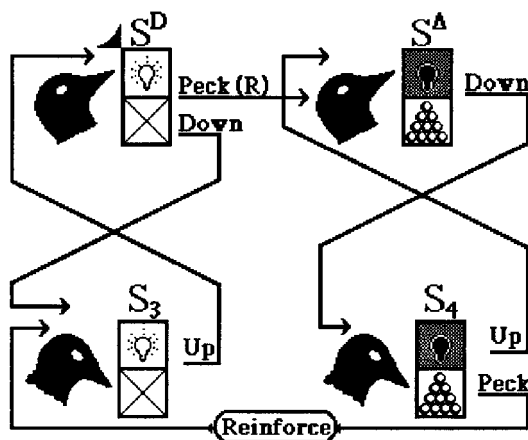


Fig. 2. Markov decision process of contingencies for CRE. Each of the four pictograms illustrates the environmental features of the corresponding state. The three features are head position (up or down), keylight status (on or off), and feeder status (open or closed). Above each state is the stimulus symbol (i.e., observation) sent to the learning model on each iteration. The four stimulus symbols (S^D , S^A , S_3 , and S_4) correspond to the four sets of discriminable environmental features. Arrows indicate state transitions due to simulated behavior. Each arrow is labeled with the deed symbol (Peck, Up, and Down) received from the learning model that produces that state transition. The peck occurring in the state that constitutes a key peck is labeled with an R for response. Arrows corresponding to reinforcer delivery (e.g., feeding) are labeled with the word "reinforce." Whenever the corresponding state transition occurs, a reinforcer symbol is sent to the learning model.

decision process (MDP; cf. Howard, 1960), and the learning model, an *adaptive controller*, forms an interactive process with the MDP. The MDP provides a computational model of the environmental contingencies and the pigeon's body. On each iteration (simulating $\frac{1}{3}$ s of real time), the stimuli produced by the simulated operant chamber are input to the controller. The controller returns calculations stipulated by the learning theory being tested, simulating the movements of the pigeon.

We use an expanded form of the Mechner diagram to specify each MDP. Figure 2 illustrates how a standard Mechner diagram (Figure 1a) is expanded for a situational simulation of CRE. This MDP consists of four *states* (the four pictures) and three *deeds* (the arrow labels). (Dynamic programming convention designates these as *actions*, but we use the term *deeds* to avoid confusion with behavioral uses of *action*.) The states correspond to ex-

ternally observable states of the environment and the pigeon's body. Deeds correspond to bodily movements. To be considered operants (e.g., key peck), the deed must occur in the appropriate state (i.e., on the top left in Figure 2).

Any of the three deeds can occur in any of the four states, resulting in a total of 12 arrows. However, for simplicity, arrows that connect a state back to itself are not shown in Figure 2 (e.g., Up deeds in the two upper states).

At the beginning of each iteration cycle, the label corresponding to the current state (either S^D , S^A , S_3 , or S_4) is output as an observation to the learning model being tested. The learning model then returns the next deed, and the testbed then determines the state transition, if any, given that deed in that state. The testbed also determines both reinforcer presentation (i.e., opening the feeder) and reinforcer delivery (i.e., feeding). Because many learning models require separate input to indicate reinforcers (e.g., Sutton & Barto, 1998), whenever the combination of deed and state (e.g., peck at an open feeder) indicates reinforcer *delivery*, the testbed outputs a "reinforcing signal" to the model. Following any state transition and reinforcer delivery, the next iteration cycle begins.

The MDP shown simplifies the model of feeder operation in that a single peck at the open feeder closes the feeder. In all the In Situ simulations presented herein, the simulated feeder actually was timed, opening for a simulated $3\frac{2}{3}$ s (11 iteration cycles). This corresponds to Ferster and Skinner's (1957, p. 19) feeder interval of between 3.5 and 4 s. During that period, any pecks at the feeder resulted in reinforcer delivery. Discussion of the modeling of timed events is deferred until Simulation 3.

Stimuli for DMOD. DMOD (Daly & Daly, 1984) is specifically designed to accept stimulus elements rather than stimulus complexes as input. To accommodate this property, each of the four stimulus complexes was translated into five separate elements (see Table 1). This divergence from the general approach of the testbed simulations demonstrates the kind of documentable flexibility available.

Table 1

Cue vectors for each stimulus and situation.

Situation	Stimulus	Cues				
		Cage	Key	Feeder	Light	Food
bof	S ₄	1	0	1	0	1
tof	S ^A	1	1	0	0	0
ble	S ₃	1	0	1	0	0
tle	S ^D	1	1	0	1	0

Note. Abbreviations for situation and stimulus values are explained in the legend for Figure 7.

Procedure

Reinforcement schedule. Within the testbed, contingencies of reinforcement are specified by the MDP. Figure 2 illustrates an MDP specifying CRF (neglecting the more usual timed opening of the feeder as noted above). For Simulations 2 and 3, we will abandon the pictograms in favor of Mechner's brackets. Because standard Mechner diagrams neglect topographic details such as the raising and lowering of the pigeon's head, we consider our diagrams (see Figures 7 and 10, presented later) to be an *expanded* form of a Mechner (1959) diagram.

Whatever type of diagram is used, the level of detail must be fine-grained enough to describe all the processes used in the situational simulation. The four states are specified in terms of three situational variables describing the status of the operant chamber (key lit or off, feeder full or empty) and the location of the pigeon's head (top, bottom). The *deed set* includes PECK (giving rise to a key peck when the pigeon's head is raised and feeding when the pigeon's head is lowered), ARCH (raising the pigeon's head), CURL (lowering the pigeon's head), and NULL. The NULL deed is included to insure that the deed set constitutes an exhaustive partition of all possible activities by the pigeon. By definition, NULL deeds model any activity that never changes the state, and the NULL arrows are not shown.

Cumulative recorder. On each cycle in which the feeder is closed, the paper roll of the cumulative recorder advances. When the learning model outputs a peck and the pigeon's head is in front of the lit key, the pen on the cumulative recorder moves up one unit. The paper roll does not advance while the feeder is open.

RESULTS

The Staddon-Zhang Model

Operant level. Under extinction prior to any conditioning, the SZ model showed bursts of responding beginning after the first 1,000 cycles. These bursts included from only a few to over 200 responses, separated by periods of no responding lasting from a few cycles to a thousand (Figure 3a). (The reader will note that the cumulative records are scaled differently from one figure to the next.) Overall, the rate of responding was approximately 0.5 responses per second.

Acquisition. Slow, steady responding was observed during CRF, with responding less than one response per 10 s (Figure 3b). This steady responding was very different from the burst-and-pause pattern observed during the operant level.

Extinction. Extinction began after about a simulated 39 min (7,000 cycles) of CRF. A rapid burst of about 50 key pecks emitted at a rate of over 2 responses per second occurred immediately. Responding then resumed the burst-and-pause pattern previously seen during the operant level. Overall rate was maintained at approximately 0.5 responses per second for the next 13,000 cycles.

Adjusting the operant level. Technical features of the SZ model limit the degree to which the likelihood of deeds can differ. With four *V* values for the four deeds, no single deed (e.g., PECK) can be emitted much less than one fourth of the time, no matter how the parameters change with conditioning. This limitation of the SZ model appears to be a major contributing factor to the unrealistically high operant level of key pecking noted above. Multiple *V* values for each deed were introduced to work around this limitation

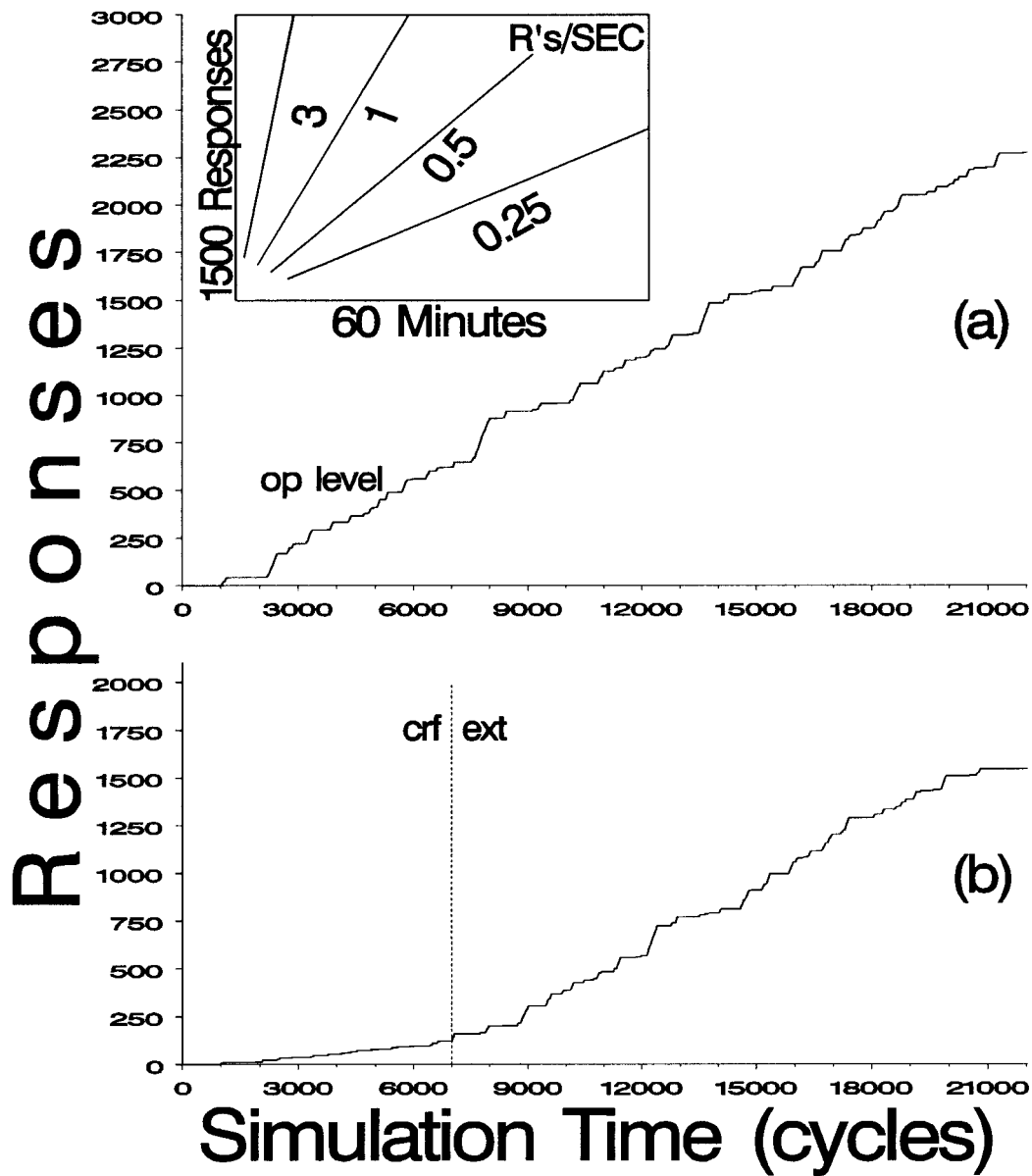


Fig. 3. Cumulative records of the SZ model under (a) EXT without prior reinforcement (operant level) and (b) CRF followed by EXT. The top graph includes two separate labeled coordinates: a traditional rate box, showing slopes for different rates of response, whose width and height indicate the scales of simulated time and responses, respectively; and a conventional x - y coordinate axis, indicating iteration cycles and simulated responses, with the 0 point denoting the start of the simulation. Vertical dotted lines indicate changes from one schedule contingency to the next.

and lower the relative observed initial rate of key pecking. Reducing this initial rate to as low as one response in 10 min in an extended session of a simulated 110 min (20,000 cycles) of CRF did not produce conditioning (not shown). Thus, within the present testbed, no

version of the SZ model was found that demonstrates reinforcement of key pecking.

The Roth-Erev Model

Operant level. Under extinction, the RE model produced a more or less steady rate of

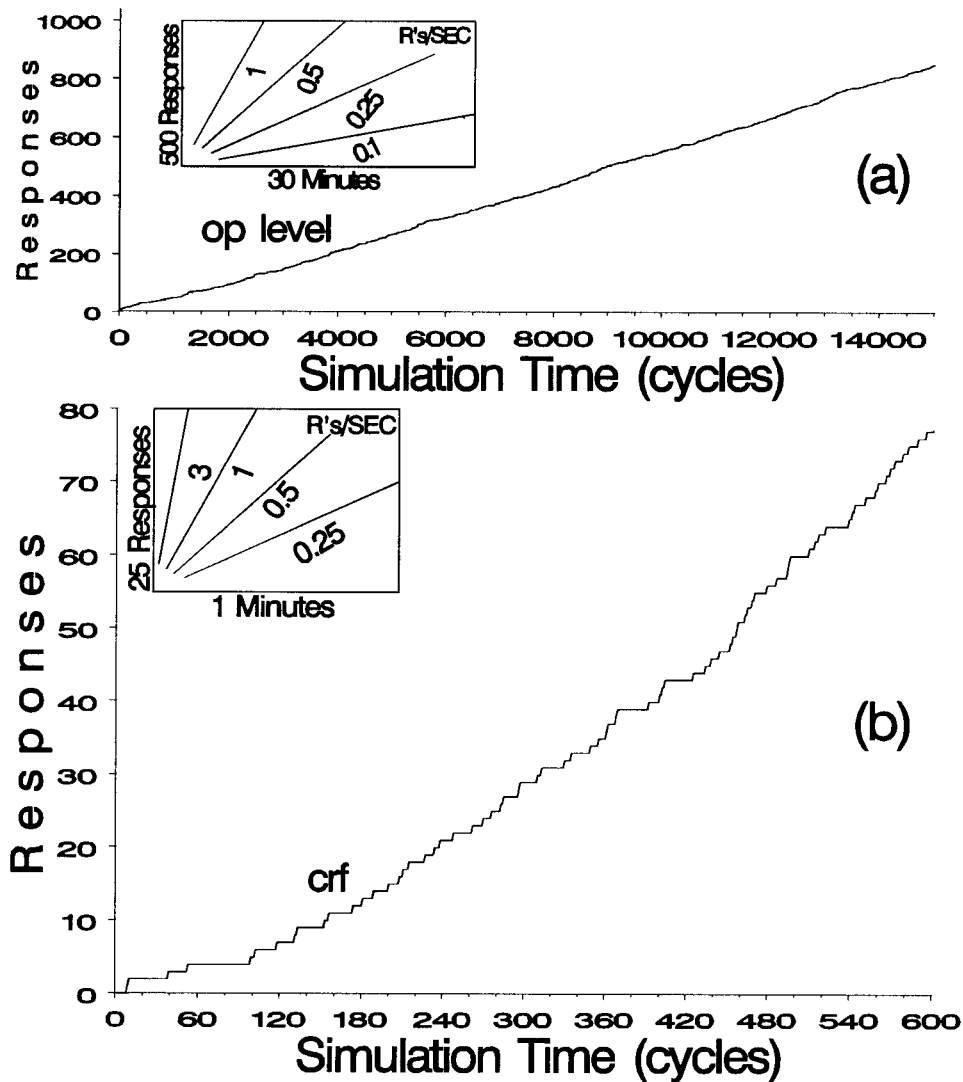


Fig. 4. Cumulative records of the RE model under (a) operant level (EXT) and (b) CRF. For CRF only, 30% of NULL deeds were reinforced on a random-ratio schedule. Note the difference in scales between top and bottom panels. See Figure 3 for an explanation of the scales and legend.

one peck per 5 s (Figure 4a). Over the length of this period, the rate slows slightly but steadily. The unrealistically high initial rate is directly due to the parameter settings chosen. These parameters were set so as to optimize the match of the cumulative record for acquisition and extinction to real animal behavior. Parameter settings that produced a more realistic (lower) operant level produced much less realistic acquisition and extinction.

Acquisition. When the initial likelihoods of the four deeds (ARCH, CURL, PECK, and

NULL) were set equal, the RE model showed rapid responding from the beginning of the CRF period. Overall rate was approximately 1 response per second, and it did not increase or decrease appreciably across training (not shown).

To slow the rate of operant responding to something more realistic, 30% of the NULL deeds were randomly reinforced (a random-ratio schedule), thereby strengthening competing responses. Figure 4b shows that key pecking increased across a simulated 3½ min

(600 iteration cycles) of CRF with this arrangement. Rate of responding increased from about 0.25 responses per second to about 0.7 responses per second. This increase demonstrates a reinforcement effect. The value of 30% was empirically selected due to extinction effects (see below).

Extinction. Rate of responding stabilized at about 1 response per second over a simulated 39 min (7,000 iteration cycles) of CRF with the conditions described above (Figure 5a). Extinction was arranged at that point, and responding showed an immediate acceleration to almost 3 responses per second. This high, steady rate was maintained over the next 15,000 cycles, with only a slight drop in response rate.

Reinforcing NULL deeds at less than 30% (not shown) lessened even this minimal drop in response rate. Higher reinforcement rates (also not shown) decreased response rates during extinction slightly, but were deemed unrealistic for reinforcement rates for competing responses.

Across parameter settings, with differently scheduled CRF and EXT components, the response rate under extinction varied inversely with the reinforcement rate for NULLs. Lower reinforcement rates for NULLs produced unrealistic results. The highest reinforcement rate deemed reasonable (30%) still failed to extinguish responding.

Adjusting the operant level. In the simulations presented thus far for the RE model, operant level was unrealistically high due to initial V values for all four deeds being set equal. In a subsequent simulation, initial V values were adjusted to produce a realistic mix of the four deeds (60 NULLs to each 4 ARCHs and 4 CURLs and to each PECK). With this mix of initial values, the RE model produced a low rate of key pecking (less than 0.05 responses per second) across the first simulated hour (11,000 cycles) of CRF. At this time, there was a sudden transition to a remarkably stable (as determined by visual inspection of the cumulative record) high rate of key pecking (over 2 responses per second) that persisted until the test was ended at a simulated 100 min (Figure 5b). The pattern of satiation typical of real pigeons (reduced responding after many reinforcers) was not seen. Further, the rate of responding was remarkably high compared to a typical pigeon (which might emit a rate of approximately 0.5 responses per second).

The Daly–Daly Model

To simulate bodily movements comparable to those of real naive pigeons in an operant chamber, initial relative likelihoods of the deeds were 360 NULLs to each 4 ARCHs and 4 CURLs and to each PECK. These initial settings were used across all DMOD* simulations.

Operant level. Figure 6a shows the DMOD* model under extinction without prior reinforcement. A low rate of pecking (0.25 responses per second) for approximately a simulated 4 min (750 cycles) was followed by a brief burst of rapid pecking (3 responses per second) followed by a return to the lower rate for a total of a simulated 18 min (3,300 cycles) before pecking ceased entirely. Despite the initial settings, the V values rapidly altered to make all deeds equally likely. This appears to be due to an increase in the avoidance component and the fact that negative V values were treated as 0.

Acquisition. Under CRF, responding began at 20 cycles and increased from less than 0.25 responses per second to around 1 response per second in just over a simulated 1 min (200 cycles). This rate of responding was maintained throughout a simulated 11 min (2,000 cycles).

Extinction. Following a simulated 6 min 40 s (1,200 iteration cycles) of CRF, responding declined over the first 200 cycles under extinction and then recovered and continued at approximately 0.4 responses per second until it abruptly ended at 5,000 cycles (see Figure 6b). This pattern provides a record that might plausibly be obtained from a real pigeon.

To assess the stability of this pattern, the procedure was repeated, but the shift to extinction was made 2 s earlier, at 1,184 cycles. A different pattern of responding during extinction was obtained in this replication (see Figure 6c). Responding was maintained fairly steadily for over 18,000 cycles without an apparent decline (about 0.4 responses per second, slowly wavering between 0.3 and 0.5 responses per second). This pattern of responding appears to be unrealistic.

From a modeling perspective, obtaining two such different patterns from such a slight change in contingencies is disconfirming, because such slight alterations do not radically af-

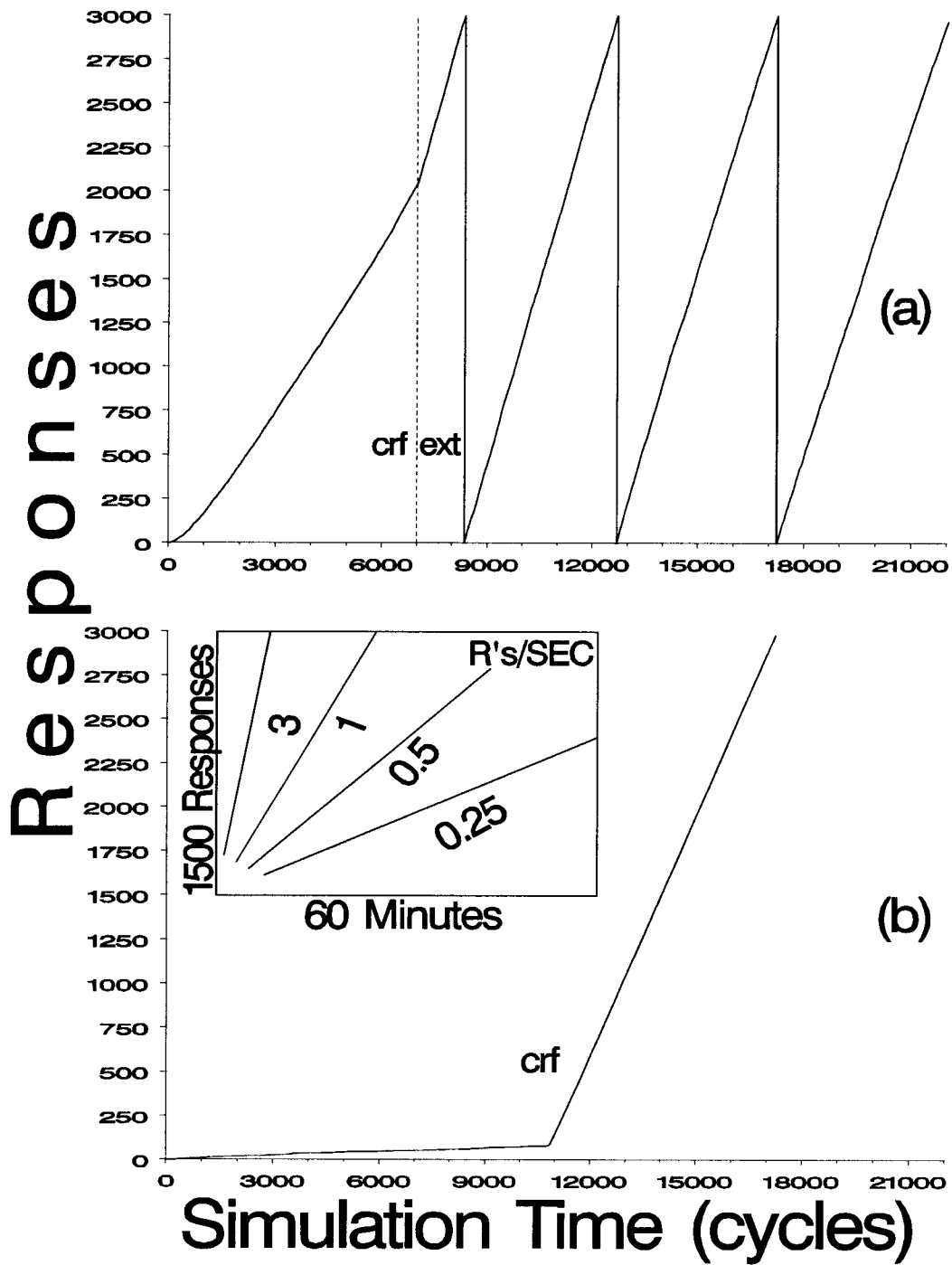


Fig. 5. Cumulative records of the RE model under (a) CRF followed by EXT and (b) CRF. The top panel (a) shows results with initial likelihoods of different body movements set equal. The bottom panel (b) shows results with more realistic initial likelihoods: few head bobs, fewer pecks, with all remaining behavior having unspecified topographies (NULL deeds). 30% of NULL deeds are reinforced on a random-ratio schedule. The response pen reset to 0 after each 3,000 responses. See Figure 3 for an explanation of the scales and legend.

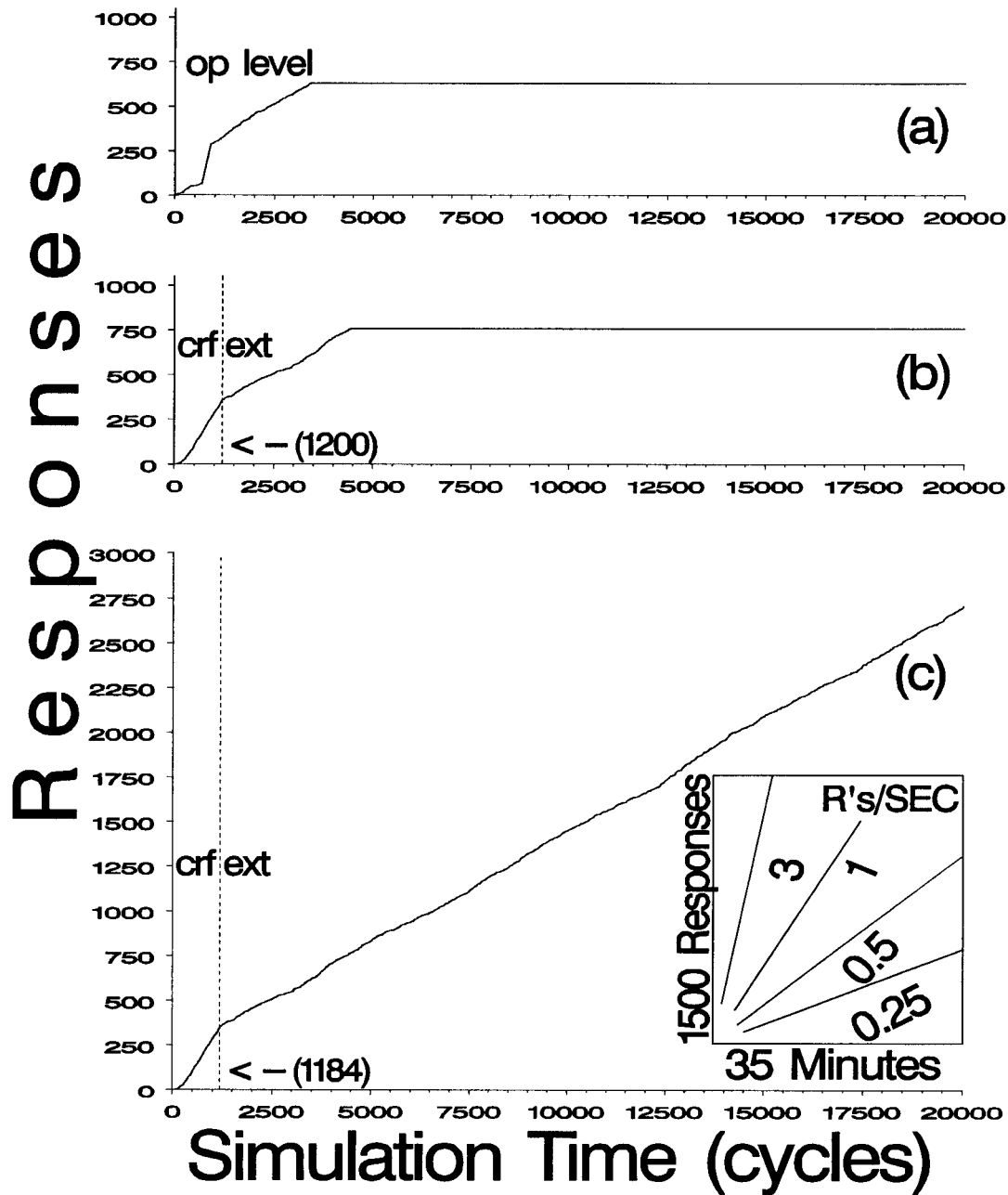
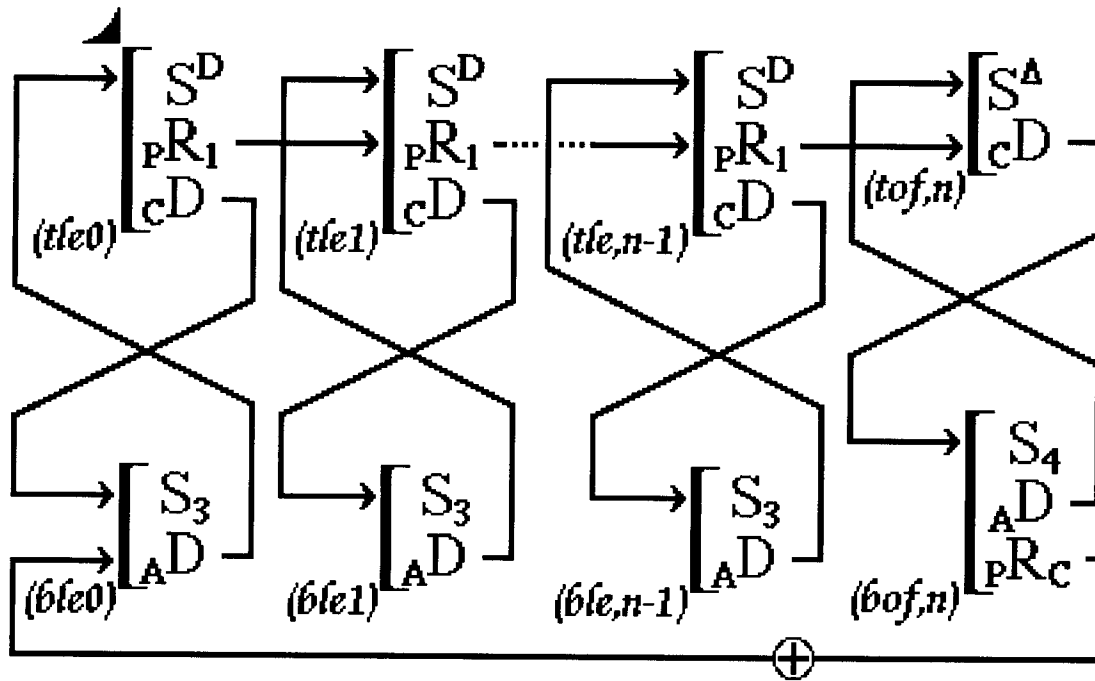


Fig. 6. Cumulative records of the DMOD* model under (a) operant level (EXT) and (b and c) CRF followed by EXT. The shift to EXT (indicated by the vertical dotted line) occurred at 1,200 cycles in the middle panel (b) and at 1,184 cycles in the bottom panel (c). See Figure 3 for an explanation of the scales and legend.

fect responding in the pigeon. From a mathematical perspective, such hypersensitivity is often found in simple nonlinear iterative systems. The dynamics of such systems are discussed under the rubrics *complexity theory* (Casti, 1997, chap. 3) and *chaos theory* (Gleick, 1987).

SIMULATION 2: FIXED-RATIO SCHEDULES

Simulating FR schedules involves consideration of a feature of MDPs called *partial observability*. The observable features in Figure



<u>Stimuli</u>	<u>(situations)</u>	<u>Responses</u>	<u>Deeds</u>	<u>Situational variable: {values}</u>
S^D lit key	(l)	R_1 keypeck	P peck	Beak location: {top=t, bottom=b}
S^A dark key	(o)	R_c consumatory	A arch	Key status: {lit=l, off=o}
S_3 closed feeder (e)			C curl	Feeder status: {full=f, empty=e}
S_4 full feeder (f)			N null	Counter (c): {0, 1, 2, 3, ... n}

Fig. 7. Expanded Mechner diagram for an FR schedule. Each bracket designates a state in the POMDP model of the FR schedule. The stimulus symbol (observation) for each state (S^D , S^A , S_3 , or S_4) is at the top of the bracket. Observations are sent to the learning model on each iteration. Every deed (bodily movement) that results in a state transition from a particular state is included within the bracket at the tail of an arrow pointing to the new state. Deeds are received from the learning model. R indicates deeds that constitute responses. D indicates all other deeds. Features distinguishing the states are identified within parentheses at the lower left of each bracket. Features are head position, keylight and feeder status, and ratio counter setting. The arrow corresponding to reinforcer delivery is labeled with a +. Brackets sharing a stimulus symbol designate states that are not discriminable from the subject's perspective.

2 (head position, keylight and feeder status) that distinguish the four states are also the only features observable under other single-key schedules, such as FR. Features unobservable from the pigeon's perspective, such as the setting of the ratio counter, are relevant to the contingencies of FR reinforcement and must be modeled. When all of the features determining the contingencies of reinforcement are observable (as in CRF), the MDP is called a completely observable Markov deci-

sion process, or COMDP. When some features are unobservable, the MDP is referred to as a partially observable Markov decision process, or POMDP (Kaelbling et al., 1998).

A POMDP is diagrammed by having more states than observation labels (see Figure 7). For FR, all of the states with the pigeon's head raised and the keylight on are labeled with observation S^D , irrespective of the ratio counter's setting. S_3 is the corresponding observation when the head is down. *Observations*

correspond to states that are discriminable by the pigeon being conditioned, which is to say, Mechner's (1959) stimulus complexes. In the present simulations, the same four stimulus complexes (S^D , S^A , S_3 , or S_4) were used as were used in Simulation 1. Figure 7 is a fully realized expansion of Figure 1b, the standard Mechner diagram for an FR schedule.

METHOD

Models and Apparatus

The SZ model was dropped from further simulations due to its poor acquisition under CRF. It was assumed that the SZ model would not produce a decelerating reinforcement rate under leaner schedules. The RE model showed better performance under acquisition. Therefore, despite poor performance under EXT, the RE model was retained. The lack of reduced responding during extinction, however, indicates that the RE model might be insensitive to differences between CRF and FR or FI. DMOD* performed well in Simulation 1, although it showed instability. It was also retained for Simulation 2.

The apparatus was identical to that of Simulation 1.

Procedure

The procedure for In Situ simulation of FR schedules is illustrated using an expanded Mechner diagram (see Figure 7). The pictograms of Figure 2 have been replaced by brackets, as in the standard Mechner diagram (Mechner, 1959). The observation labels are placed inside the brackets at the top (Millenson, 1967). In most other respects, this expanded Mechner diagram works similarly to the pictographic diagram: As before, arrows for deeds that do not change the situation are not shown but are still recorded as simulated time on the cumulative record as long as the feeder is closed. This expanded Mechner diagram also does not indicate that the feeder was timed to open for 11 iteration cycles.

The strategy for testing the models under FR was to retain the best parameters found for each model under CRF/EXT and to subject each model first to CRF until stable responding (as determined by visual inspection) was established, followed by a sequence of FR 2, FR 4, FR 8, FR 16, and FR 32.

RESULTS

The Roth-Erev Model

The data in Figure 8 show that the RE model achieved stable responding by a simulated 5.5 min (1,000 iteration cycles). Each of the subsequent ratios was presented for that same amount of time. As is standard with cumulative records for schedules other than CRF, whenever the feeder opened, a deflection of the response pen (pip) was added to the cumulative record in this and the subsequent figures.

Performance under FR 1 was, of course, identical to the performance previously shown for CRF: steady responding at about 1 response per second. With the shift to FR 2, response rates increased from 1 response per second to 2.5 responses per second. Each successive shift to a higher FR value shifted responding to a higher rate. Responding was not, however, quite as stable as it was during the CRF schedule. Periods of more rapid and less rapid responding alternated within a narrow range (perhaps 2 to 3 responses per second). No interreinforcer pattern of pausing and responding emerged. There was no tendency toward break-and-run, for example, during the 91 reinforcers obtained under FR 32.

The Daly-Daly Model

The DMOD* model achieved stable responding after less than a simulated 7 min (1,200 iteration cycles). Each of the subsequent ratios was presented for that same amount of time (see Figure 9a).

Performance under FR 1 was identical to the previous CRF performance: steady responding just above 1 response per second. Performance under the sequence of FR schedules was as follows: Responding was somewhat slowed during FR 2 and became more variable in rate. There was, however, no long-term decline during this 1,200-cycle period. FR 4 through FR 16 continued to maintain pecking at this somewhat slower (0.4 responses per second) overall rate. As with the RE model, no apparent pattern of responding developed between reinforcers (i.e., no break-and-run pattern). Continued training at FR 32 (55 reinforcers) showed responding at the same overall rate, with occasional hints

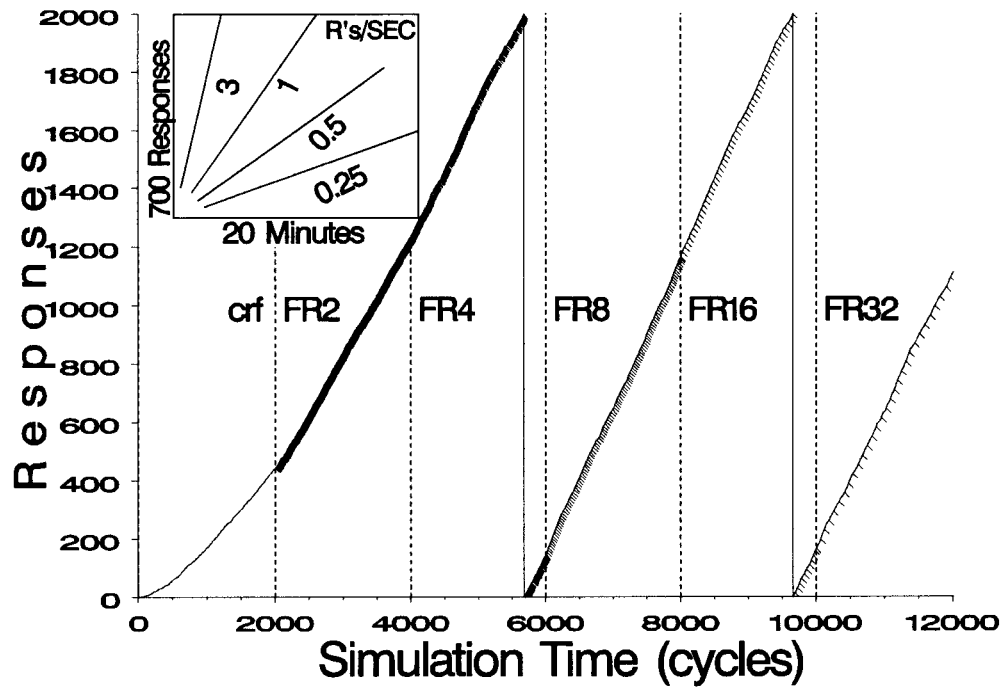


Fig. 8. Cumulative record of the RE model under CRF followed by FR 2, FR 4, FR 8, FR 16, and FR 32. Vertical dotted lines indicate ratio changes. Pips indicate reinforcer delivery. The response pen reset to 0 after each 1,500 responses. See Figure 3 for an explanation of the scales and legend.

of a pause developing early in some inter-reinforcer intervals.

Figure 9b shows a replication of DMOD* under control of 1,200 iteration cycles of CRF, followed by a simulated 2 hr 40 min (over 28,000 cycles) of FR 32. The direct shift from CRF to FR 32 produced a pattern of responding similar to that generated by the progressive approach to FR 32. Overall rate was 0.4 responses per second, and the inter-reinforcer pattern varied from one reinforcer to the next. This more extended training on FR 32, however (approximately 132 reinforcers), did not produce a clear development of the characteristic FR break-and-run pattern of responding (Ferster & Skinner, 1957) between reinforcers.

DISCUSSION

In Situ simulations of ratio schedules challenged computational learning theories in ways that more conventional simulations cannot. Trial-by-trial simulations, for instance, present all relevant stimulus information to the model on each iteration. Partial observability allows the In Situ testbed to evaluate

learning models with respect to important features of real animal behavior. To the degree that a pigeon or a learning model responds differently to a lit key shortly after reinforcement than it does to the same lit key after several key pecks, that pigeon is under control of the situation rather than merely the momentary stimulus.

The failure of either model to develop a postreinforcement pause indicates a failure of those models to address Skinner's question: "What are the relevant features of the environment, and how are they to be measured and controlled?" (1984a, p. 514, 1988a, p. 83). Traditionally, learning theories are expressed in terms of generic stimuli and responses, neglecting distinctions such as that between the momentary stimulus and the total ongoing situation. By explicitly modeling these distinctions, the In Situ testbed forces theories to demonstrate the ability to come under the control of temporally extended phenomena despite having been exposed to the stimuli on a moment-to-moment basis only. Thus, the failures of these models demonstrate important capabilities of the In Situ

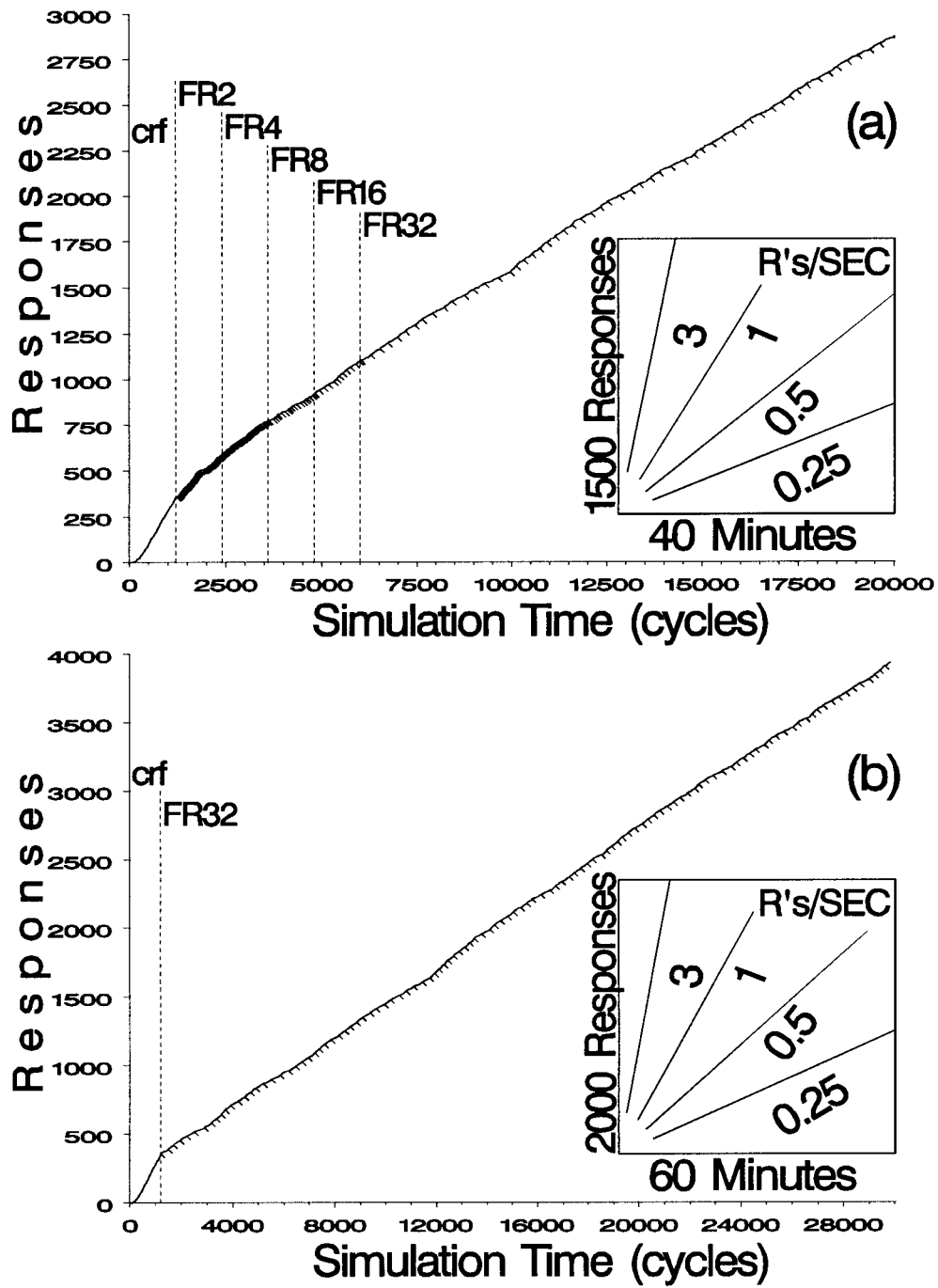


Fig. 9. Cumulative records of the DMOD* model under FR schedules. The top panel (a) shows the DMOD* model under CRF followed by FR 2, FR 4, FR 8, FR 16, and FR 32. The bottom panel (b) shows the DMOD* model under CRF followed immediately by FR 32. Vertical dotted lines indicate ratio changes. Pips indicate reinforcer delivery. Note the difference in scales between top and bottom panels. See Figure 3 for an explanation of the scales and legend.

testbed as well as shortcomings of the theories themselves.

SIMULATION 3: FIXED-INTERVAL SCHEDULES

Computational models of reinforcement schedules face a special problem in that ratio and interval schedules treat time differently: It can be an independent variable in interval schedules but not in ratio schedules. Computational systems use states to model independent variables. Markovian and non-Markovian models, which support different types of mathematical analysis, place different restrictions on the relations between states and time. For a Markovian model, such as those used here, the value of every independent variable must be available as a feature of the current state; history does not matter. Markovian models of interval schedules, in which time is an independent variable, are thus problematic.

Models of learning are inherently non-Markovian, in that history matters a great deal. Metaphors such as "memory" are invoked to suggest that the brain somehow translates history into the current "state" of the brain. The problem of modeling environments, such as is done with the In Situ testbed, is nowhere near so complex. The operant chamber itself is a Markovian device, using timers to encode the intervals elapsed since various past events as current states of the timing mechanism. The "state" of the real operant chamber includes the then-current timer setting. Being restricted to modeling only external environmental and body events, the In Situ testbed can resolve the problem of modeling interval schedules in exactly the same fashion as is done by the operant chamber. The setting of a simulated timer is included as a feature of the state definition, as shown in Figure 10.

METHOD

Models and Apparatus

Both the RE and DMOD* models were retained for the FI simulation.

The apparatus was identical to that used in the CRF and FR simulations.

Procedure

The procedure for In Situ simulation of FI schedules is illustrated in Figure 10, which is an expanded version of the standard Mechner diagram (cf. Figure 1c). Standard Mechner diagrams use distinct symbols (e.g., the T in Figure 1c) for state transitions initiated by timers in the operant chamber as opposed to responses by the organism (Mechner, 1959). In our expanded version, any transition events not initiated by the behavior of the organism are designated with an H subscripted by a function giving the definition of the event. The timer setting is included among the situational variables. Within the simulation, each iteration within an interval constitutes a new state. Thus, each bracket summarizes the many states with timer values satisfying the inequality listed within the parentheses. FI schedules, like FR schedules, are partially observable, in that the timer setting, like the ratio counter, is not observable by the pigeon. For simplicity, we continue to represent the closing of the feeder with an R rather than with an H subscripted to indicate the $3\frac{2}{3}$ -s feeder interval.

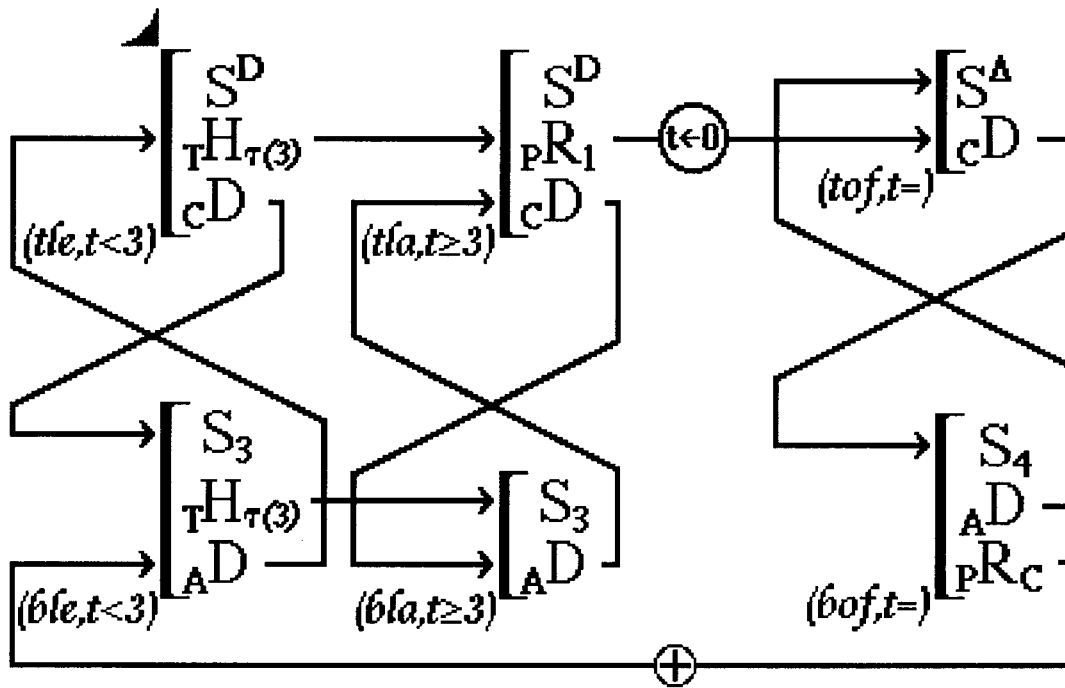
The strategy for testing the models under FI was to retain the best parameters found for each model under CRF/EXT and to subject each model first to CRF until stable responding (as determined by visual inspection) was established, followed by an FI 1-min schedule.

RESULTS

The Roth-Erev Model

Figure 11 shows the RE model under control of a simulated 11 min (2,000 iteration cycles) of CRF (FR 1), followed by a simulated 16 min (almost 3,000 cycles) of FI 1 min. Performance under FR 1 was the same as that shown previously: steady responding at about 1 response per second, stable after 2,000 iteration cycles.

Under the FI 1-min schedule, responding increased with the change from CRF to FI (as it had for extinction). Responding was maintained at a fairly steady pace of about 2 responses per second throughout the 13 reinforcers provided according to this schedule. Although there were minor moment-to-moment variations in rate of responding, there was no development of a pattern of responding paced by the interreinforcer intervals.



Legend as per Figure 7, with the addition of:

Hap	Situational variable: {values}
$\tau(3)t > 3:00$	Timer (t): {00:00, ... 59:59}

Fig. 10. Expanded Mechner diagram for the FI 3-min schedule. Each bracket designates a state as distinguished by four features: head position, keylight and feeder status, and a range of values for the interval timer. In addition to the deeds used in Figure 7, state transitions can also occur as the result of events (called haps) occurring in the feeder mechanism. These are usually timed events and are designated by the letter H (equivalent to a T in a standard Mechner diagram). The condition triggering the event is indicated in the subscript $\tau(3)$. The additional feeder status (designated with an a) indicates that the feeder will open on the next key peck. See Figure 7 for an explanation of the other abbreviations and symbols.

Neither extinction-like nor scalloped patterns appeared.

The Daly-Daly Model

Figure 12 shows responding generated by the DMOD* model under control of less than simulated 7 min (1,200 iteration cycles) of CRF, followed by a simulated 1 hr 40 min (over 18,000 cycles) of FI 1 min. Performance under CRF was identical to that previously shown: steady responding just above 1 response per second that stabilized after 1,200 iteration cycles.

Under the FI 1-min schedule, as with extinction, responding declined across the first interreinforcer interval. Responding then recovered and was maintained at an overall

rate of 0.4 responses per second for the duration of training (approximately a simulated 105 min). There was no clear development of a response pattern filling the interreinforcer intervals. Some intervals provided sharp run-and-break patterns (reminiscent of the mini-extinction curves sometimes seen early in the development of FI patterning). A few intervals appeared to start with a lower rate and end with a higher rate (scallop pattern), but this pattern did not increase across training.

When this training was extended for a period simulating almost 2 hr 40 min (over 28,000 cycles) of FI 1 min, again, scallop patterns were not observed to become more frequent with training (not shown).

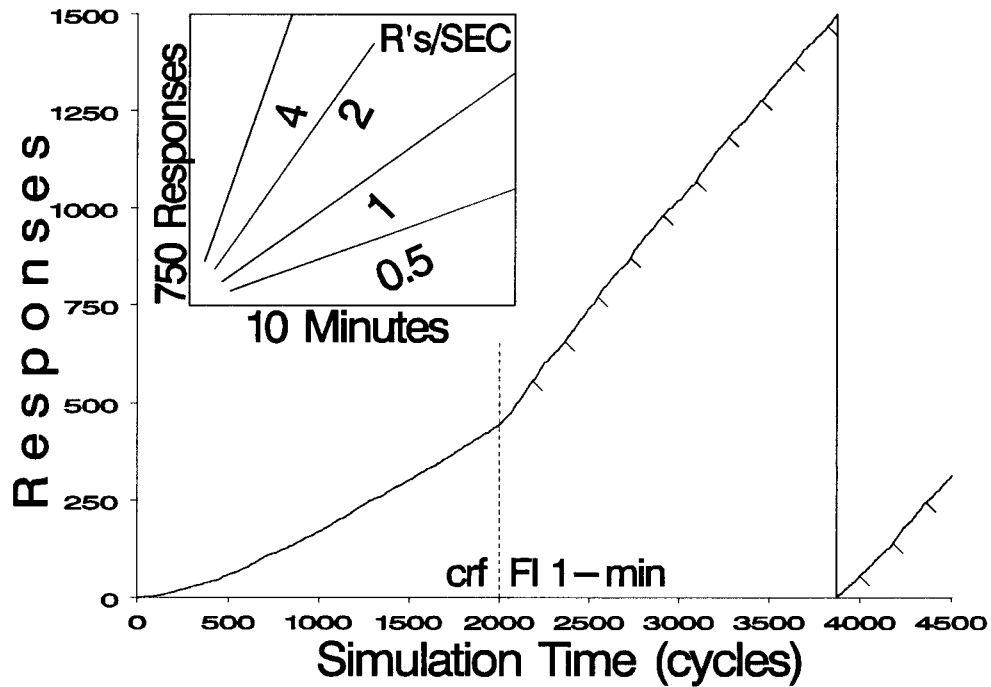


Fig. 11. Cumulative record of the RE model under CRF followed by FI 1 min. Vertical dotted lines indicate the introduction of the FI schedule. Pips indicate reinforcer delivery. The response pen reset to 0 after each 1,500 responses. See Figure 3 for an explanation of the scales and legend.

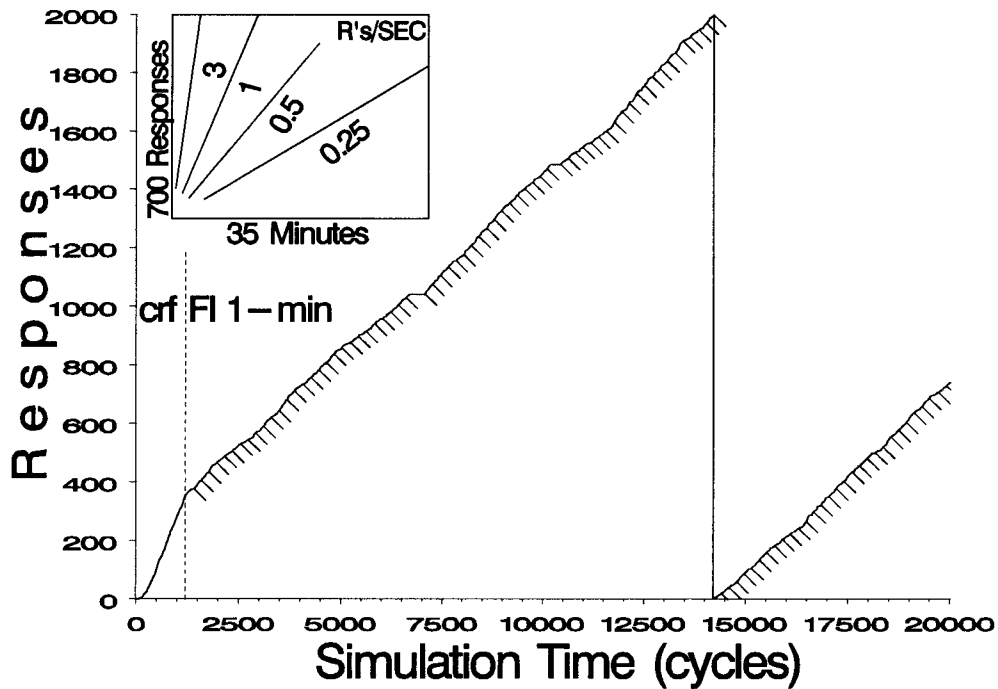


Fig. 12. Cumulative record of the DMOD* model under CRF followed by FI 1 min. Vertical dotted lines indicate the introduction of the FI schedule. Pips indicate reinforcer delivery. The response pen reset to 0 after each 1,000 responses. See Figure 3 for an explanation of the scales and legend.

DISCUSSION

Simulations of FI schedules with the In Situ testbed demonstrated the inability of both the RE and DMOD* models to simulate pigeon responding under those schedules as well as the capability of the In Situ testbed to accurately generate both interval and ratio schedule performances.

GENERAL DISCUSSION

We have presented example results from a novel testing procedure. Three questions are considered here: How well did the procedure work? Of what value is the procedure to contemporary behavior analysis? What conceptual issues arise in considering adoption of the procedure?

Before we consider these questions, however, it should be noted that our goals for the present report are aspirational. We wish to demonstrate the potential of In Situ rather than to reject models here. It is hardly a mark against the Staddon–Zhang model that it was unable to generate organized temporal sequences of responses. The Staddon–Zhang model was designed to emphasize a particular point, not to simulate the changes in behavior of a whole pigeon. It would be unreasonable, therefore, for us to claim that the present study has falsified any of the three models tested. Examined on their own terms, all three models performed well in the prior publications designed to show their strengths. Examined on the novel terms proposed by Church (1997) and by ourselves, however, all three performed poorly. With this report, we are advocating that modelers take up the challenge of creating new behavioral models that are powerful enough to handle simulations like the In Situ simulations presented herein.

TESTBED ASSESSMENT

Evaluating any testing procedure involves determining reliability and validity (Anastasi, 1988; Carmines & Zeller, 1979).

Reliability

A testing procedure is reliable to the degree that the results limit the amount of random (unbiased) error or “noise” (Carmines & Zeller, 1979, pp. 13–15). The underlying

strategy for any reliability assessment is to repeat measures under conditions in which only uncontrolled factors can alter the result. The degree to which the results vary across these measurements determines the assessed reliability.

Test-retest. In the usual psychological testing, test-retest procedures form the core of reliability assessment. Examinees and procedures are held constant; only occasions are allowed to vary (Shavelson, Webb, & Rowley, 1989). In the present context, in which one computer program is being used to test another, simply running the same computer routines (with the same random seed) on a second occasion will generate identical results. Therefore, the notion of merely repeating a measurement is not applicable.

Instead, to determine reliability, parameters of the testbed that model factors that are usually uncontrolled in real operant conditioning experiments can be systematically varied across simulations. For example, two factors over which real operant procedures are generally reliable are the choice of individual subject (In Situ represents an individual by selecting a unique pseudorandom seed for the random number generator) and small variations in the duration of individual components or events in a sequence of schedules (the precise number of iteration cycles that a particular MDP is in effect during the simulation). As noted in Simulation 1, only in the case of DMOD* did this type of variation produce any effect. The failure of these manipulations for our implementation of the DMOD model challenges either the stability of DMOD* or the reliability of the In Situ testbed. The fact that only this one model exhibited these difficulties suggests that either DMOD* has some instability or the In Situ testbed is an unreliable procedure for evaluating DMOD*, despite being a reliable procedure for the other two models. In our discussion of construct validity below, we will argue for the former.

Internal consistency. A procedure is said to be internally consistent if task difficulty and examinee attributes are systematically related. For this reason, tests of internal consistency are often considered to be measures of validity as well as reliability. An important advantage to the use of Mechner’s (1959) classification system for reinforcement schedules is

that it provides an external measure of complexity and, hence, of task difficulty for different reinforcement schedules. The greater the number of situations and deeds, the fewer paths through the expanded Mechner diagram result in reinforcement. From this perspective, CRF is objectively “simpler” than FR or FI, and higher ratios and longer intervals are more “difficult” than lower and shorter ones, respectively.

One piece of evidence that the In Situ testbed exhibits internal consistency is that none of the three models failed to simulate animal behavior for a simpler schedule and then succeeded with a more complex one. Only the one model that adequately simulated behavior under both CRF and EXT, DMOD*, managed even a rough simulation of the various ratio and interval schedules.

Informativeness. Although not strictly a component of either reliability or validity alone, an important feature of any test is that it be informative. A test so easy that everyone gets an A and a test so difficult that everyone gets an F are both equally uninformative for purposes of comparing students. Because a principal purpose of the In Situ testbed is model comparison, informativeness is vital. Of the three models tested, one failed the first and simplest simulation, one managed acquisition under CRF but failed to show realistic extinction, and the last passed Simulation 1, but failed to show a break-and-run pattern under FR or an FI scallop in Simulations 2 and 3. Therefore, the In Situ testbed was informative in distinguishing among the three models. As improved theories become available, the entire array of reinforcement schedules used in over 40 years of published data are available to provide simulations of varying difficulty, maintaining informativeness without risking internal consistency.

Validity

A testing procedure is valid to the degree that the results limit the amount of nonrandom (biased) error (Carmines & Zeller, 1979, pp. 13–15). To the degree that a test measures what it is designed to measure, it is valid. The central challenge in assessing validity is in determining the “true” value of what is being measured independently of the test used to perform that measurement.

Content validity. A test item exhibits content

validity if it identifies some aspect of the environment that controls the behavioral competence to be tested (Anastasi, 1988). For the In Situ testbed, examination of its logic by researchers who are familiar with the mechanical details of the operant chamber is the best test of content validity. Because the logic of the testbed is presented in the form of expanded Mechner diagrams, the evidence of content validity is made explicit and public in a form that does not require the reader to be an expert in computers or computer programming. To the degree that researchers are convinced that each of these expanded Mechner diagrams accurately identifies the contingencies instantiated by the corresponding reinforcement schedule, the In Situ testbed can claim content validity. Thus, we leave the assessment of content validity to the reader.

Criterion validity. A procedure is said to exhibit criterion validity if some independent evaluation of the examinees corresponds with the evaluation made by the procedure being assessed. We have two independent measures of the three models examined. The relative number and complexity of phenomena successfully addressed in the published reports of the three models provide a rough measure of their relative efficacy. The relative number of internal components (equations and parameters) in the three models provides a rough measure of their potential explanatory power. In each case the order is the same. The DMOD model is more complex than, and has been demonstrated to account for more phenomena than, the Roth–Erev model, which is more complex than, and has been demonstrated to account for more phenomena than, the Staddon–Zhang model. The results of In Situ testing produce the same order, demonstrating a degree of criterion validity.

Construct validity. In the context of testing humans, the notion of construct validity is difficult for behaviorists to accept, because construct validity is often described in terms of traits—traits supposedly possessed by the examinee. In the present context, however, in which the examinee is a set of equations, no such problem arises. What might be called our examinees’ complete anatomy and physiology are available for inspection, and no surplus meaning is implied. To the degree

that the In Situ test results, expressed in terms of the phenomena of operant learning—such as success (or failure) in acquisition or extinction, evidence (or lack thereof) of an FI scallop or ratio strain, and so on—can be related to mathematical features of the model being tested (the model's internal traits, so to speak), the testbed exhibits construct validity.

It is possible that the In Situ testbed might have been reliable and valid in every other respect and still have lacked construct validity. The testbed might deliver accurate, reliable results determining which models succeeded or failed in simulating various operant phenomena under control of various reinforcement schedules, but it is only if these results guide the theoretician to specific weaknesses or lacunae of the theory that the testbed exhibits construct validity.

The results of the three simulations show that the In Situ testbed has construct validity. For example, following the simulations, we are in a position to recommend adjustments in the theories that would correct weaknesses exhibited in those simulations. The test of the SZ model pointed to the lack of a third parameter for setting the operant level independent of the conditionability, thus confirming the original authors' position on this point (Staddon & Zhang, 1991, p. 289). The test of the RE model illustrated the precise weaknesses that led the original authors to advance a second, more complex model (Erev & Roth, 1998; Roth & Erev, 1995). This second model has three additional parameters, including a "forgetting" parameter to account for extinction. The test of DMOD* confirmed the original authors' admission that their model lacked a specific approach to "performance mapping" (Daly & Daly, 1982, p. 447). Further, the test of DMOD* also identified a second, hitherto unidentified problem: There is no means of identifying which responses are to be used in determining the magnitude of "secondary" reinforcement. Finally, the In Situ test of DMOD* detected hypersensitivity of the results to both the pseudorandom seed (data not shown) and the precise duration of CRF preceding EXT. This hypersensitivity strongly suggests that nonlinearities in the model generate chaotic (or at least catastrophic) fluctuations. In each and every case, the results

from the In Situ testbed were highly diagnostic. Presumably, after these tests, the theoreticians are in a better position to improve their theories.

That we hold back from suggesting specific solutions to the problems identified indicates an ineluctable awkwardness in testing someone else's theory. When the tests revealed an inadequacy in a model, we felt obligated to restrict the extent to which we modified the model to help it pass the test. But, the question remains, were we assiduous enough? Were we persistent enough? Were we creative enough? Did we give each model every opportunity to pass these difficult tests? Did we represent the theories well enough? Our purpose in the present article is to demonstrate, insofar as possible, situational tests of each theory as published. We hope the authors of the theories will use the diagnoses provided by these simulations to generate improvements in their theories. We hope that theoreticians in general will make use of the In Situ testbed to evaluate and improve their emerging theories of learning.

Our ultimate goal is to stimulate the generation of new and better theories by helping theoreticians test their own models using the In Situ testbed. We do not wish to test theories ourselves, but we can provide a set of standardized statistical routines for such testing.

THE ROLE OF DIRECT ANALYSIS

We propose that these situational analyses are a form of direct analysis. Typically, the operant literature provides a functional analysis, showing the functional relation between variables (e.g., Shull, 1991). Relations are demonstrated between conditions of the experiment and data from an experimental analyses of behavior, carried out at either a molar or a molecular level.⁴ Direct analysis is not a form of and cannot replace functional analysis. Direct analysis does not produce either an experimental analysis of behavior or a functional analysis. Instead, direct analysis can be used *in concert with* a functional analysis to further the analysis of behavior.

⁴This distinction (Baum, 1973; Meazzini & Ricci, 1986) is often made to contrast accounts describing overall functional relations (molar) versus moment-to-moment (molecular) changes in behavior. The present method is intended to be useful to both approaches.

Defining Direct Analysis

Direct analysis uses mechanical procedures to generate behavioral predictions from hypotheses given in the form of models or theories. The key to direct analysis is that the level of granularity of the mechanically generated predictions is set finely enough to match the level of measurement for the actual behavior emitted by the organism in the laboratory (Kemp & Eckerman, 1995). One test of this match is that the structure of the data file produced by a direct analysis should be identical to the structure of the data file obtained from an experiment (Church, 1997). (The only difference is that the direct analysis may produce a record of events that cannot be recorded in an experiment for practical reasons.) Thus, predictions generated by a procedure for a direct analysis can be compared *directly* to the results obtained from an experimental analysis of behavior, regardless of how fine a level of detail is specified for the experimental analysis.

The present study illustrates direct analysis by generating predictions from computational models of learning at a level of granularity fine enough that the predictions are displayed as cumulative records. Ordinarily, neural networks and other computational models of learning do not (and we argue cannot) generate predictions at that level of granularity. Therefore, they must be analyzed *indirectly*, whether the comparison is with predictions made at either a molar or a molecular level. For example, neural networks are usually evaluated by comparison with other theories, and are not compared directly to data. Similarly, machine-learning models are usually expressed at the abstract level of symbolic artificial intelligence, in terms of problems and solutions rather than in terms of behavior. We advocate the use of direct analysis to augment extant indirect analyses.

Integrating Direct and Functional Analyses: An Example

To illustrate how direct analysis can be integrated with and can augment a functional analysis, we will consider Shull's (1979) functional analyses of the postreinforcement pause. Shull proposed that the postreinforcement pause consists of two components: (a) a true pause consisting of nonterminal be-

havior primarily controlled by reinforcers that are implicit in the situation and are not provided by the experimenter and (b) unmeasured terminal behavior that is precurrent to the first postreinforcement response and is putatively also controlled by the experimenter-provided reinforcers. Shull offered a number of arguments, supported by functional analyses, in support of this thesis.

Suppose a computational theory (or theories) were offered to account for FI patterning as described by Shull's (1979) data. Suppose further that the theory were tested in the In Situ testbed and showed good acquisition and extinction of the operant, as well as reasonable overall rates of responding under both FI and FR reinforcement schedules. Suppose, however, that the theory *failed* to show a postreinforcement pause. (Such models are not out of the question. Erev & Roth's, 1998, extension of the RE model shows better extinction performance than did RE and is therefore a candidate. Ideally, the instability found in DMOD* can be rectified.) This direct analysis would then be able to provide a basis for further testing of computational models representing Shull's hypothesis using the In Situ testbed.

The feature of the In Situ testbed that enables it to augment and support a functional analysis such as Shull's is its flexibility in simulating postulated contingencies and deeds at whatever level of detail the functional analysis demands. Shull (1979, pp. 214–215; Capehart, Eckerman, Guilkey, & Shull, 1980), for example, offered a functional analysis of how the amount of terminal behavior affects the terminations-per-opportunity function. Such behavior could be modeled with the In Situ testbed in a straightforward manner: Treat the sequence of deeds, beginning with termination of the nonterminal behavior (e.g., returning the head to face the panel) and ending with the head positioned in front of the lit key, as a sequence of *stages* (in fact, proposed by Shull). In short, the model being tested, under control of the chosen schedule, would generate differing amounts of nonterminal and terminal behavior. Features of the testbed can be varied so as to model the effects of the specific variables postulated in Shull's analyses. Further, these features could be varied identically for all models under test so as to aid comparisons.

A central feature of Shull's functional analysis is that less control over the time spent in unmeasured terminal behavior is exercised by FI than by FR schedules. This hypothesis can be tested for various models using direct analysis with the In Situ testbed. Further, the contribution of the various components suggested by Shull (e.g., number of stages, duration of each stage, number of types of non-terminal behavior, amount of nonterminal reinforcement available, etc.) can be systematically assessed.

In fact, direct analysis allows a researcher to test hypotheses, at any point on the molar to molecular continuum, by simulating those sources and examining the resultant effects in a form of data directly comparable to that used to report the original effects in the empirical literature. Direct analysis provides a bridge between functional analyses and the experimental analyses of behavior.

CONCEPTUAL ISSUES

Further discussions of the issues raised above are not appropriate for a method-focused report such as this. Yet, we would be remiss if we avoided indicating relevant issues for future discussions as well as resources available for the reader to investigate. For instance, issues surrounding the incommensurability of bodily movements and accomplishments that lead to the neural network dilemma are discussed in Weiss (1924, pp. 42–44, 1925, pp. 55–56), Guthrie (1940), Wittgenstein (1953, §612), Austin (1962, p. 112), Goldman (1970), Lee (1981, 1983, 1986, 1988, 1996), and Kemp (1996). Techniques for evaluating the microstructure of behavior are addressed by Mechner (1992), Lee (1996), and Kemp and Eckerman (1995). Mechner diagrams are presented in Mechner (1959), Weingarten and Mechner (1966), and Millenson (1967). Markov decision processes and partially observable Markov decision processes, the mathematical formalisms underlying the expanded Mechner diagrams, are presented in Kaelbling et al. (1998), Howard (1960), Sondik (1971), and Monahan (1982). In addition, relations between these formal devices and computational models of behavior (called reinforcement learning) are discussed by Barto, Bradtke, and Singh (1995), Singh et al. (1994), Jaakola et al. (1995), and Sutton and Barto (1998).

Issues surrounding the computational modeling of behavior as environment–behavior interactions, discussed under the rubric of situativity theory, are presented in Clancey (1993, 1997), D. A. Norman (1993), Vera and Simon (1993a, 1993b), and Suchman (1993). Situational semantics, the formal underpinnings to situativity theory, are discussed by Barwise and Perry (1983) and Burke (1991), among others. The variant of information theory necessary to situational semantics and the present analysis was authored by Dretske (1981), and the implications are discussed by Kemp and Eckerman (1995). General non-technical introductions to the issues surrounding the computational simulation of interacting systems are given in Casti (1997) and Gleick (1987).

Issues in the evaluation of computational models of behavior have been largely neglected. Exceptions are Abelson (1968), Newell and Simon (1972), Einhorn, Kleinmuntz, and Kleinmuntz (1979), Staddon (1988), Massaro (1988), Kemp and Eckerman (1995), and Church (1997). Many issues involving the integration of these topics with the study of learning remain, as does the issue of their potential impact on behavior analysis.

REFERENCES



- Abelson, R. P. (1968). Simulation of social behavior. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology: Vol. 2. Research methods* (2nd ed., pp. 274–356). Reading, MA: Addison-Wesley.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: MacMillan.
- Austin, J. L. (1962). *How to do things with words* (2nd ed., J. O. Urmson and M. Sbisà, Eds.). Cambridge, MA: Harvard University Press.
- Barto, A. G., Bradtke, S. J., & Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72, 81–138.
- Barwise, J. (1989). *The situation in logic* (CSLI Lecture Notes No. 17). Stanford, CA: CSLI.
- Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge, MA: MIT Press.
- Baum, W. M. (1973). The correlation-based law of effect. *Journal of the Experimental Analysis of Behavior*, 20, 137–153.
- Baum, W. M., Schwendiman, J. W., & Bell, K. E. (1999). Choice, contingency discrimination, and foraging theory. *Journal of the Experimental Analysis of Behavior*, 71, 355–373.
- Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.
- Burke, T. (1991). Peirce on truth and partiality. In J.

- Barwise, J. M. Gawron, G. Plotkin, & S. Tutiya (Eds.), *Situation theory and its applications* (Vol. 2, pp. 115–146). Stanford, CA: CSLI.
- Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 58, 313–323.
- Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.
- Capehart, G. W., Eckerman, D. A., Guilkey, M., & Shull, R. L. (1980). A comparison of ratio and interval reinforcement schedules with comparable interreinforcement times. *Journal of the Experimental Analysis of Behavior*, 34, 61–76.
- Carmine, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Quantitative Applications in the Social Sciences, No. 17). Newbury Park, CA: Sage.
- Casti, J. L. (1997). *Would-be worlds*. New York: Wiley.
- Church, R. M. (1997). Quantitative models of animal learning and cognition. *Journal of Experimental Psychology: Animal Behavior Processes*, 23, 379–389.
- Clancey, W. J. (1993). Situated action: A neurophysiological interpretation response to Vera and Simon. *Cognitive Science*, 17, 87–116.
- Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge: Cambridge University Press.
- Daly, H. B., & Daly, J. T. (1982). A mathematical model of reward and aversive nonreward: Its application in over 30 appetitive learning situations. *Journal of Experimental Psychology: General*, 111, 441–480.
- Daly, H. B., & Daly, J. T. (1984). DMOD—A mathematical model of reward and aversive nonreward in appetitive learning situations: Program and instruction manual. *Behavior Research Methods, Instruments, & Computers*, 16, 38–52.
- Donahoe, J. W., & Palmer, D. C. (1994). *Learning and complex behavior*. Boston: Allyn & Bacon.
- Donahoe, J. W., Palmer, D. C., & Burgos, J. E. (1997). The S-R issue: Its status in behavior analysis and in Donahoe and Palmer's *Learning and Complex Behavior*. *Journal of the Experimental Analysis of Behavior*, 67, 193–211.
- Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
- Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgement. *Psychological Review*, 86, 465–485.
- Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88, 848–881.
- Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York: Appleton-Century-Crofts.
- Gleick, J. (1987). *Chaos: Making a new science*. New York: Viking.
- Goldman, A. I. (1970). *A theory of human action*. Englewood Cliffs, NJ: Prentice Hall.
- Guthrie, E. R. (1940). Association and the law of effect. *Psychological Review*, 47, 127–148.
- Herrnstein, R. J. (1997). *The matching law: Papers in psychology and economics* (H. Rachlin & D. I. Laibson, Eds.). Cambridge, MA: Harvard University Press.
- Howard, R. A. (1960). *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press.
- Jaakola, T., Singh, S. P., & Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable Markov decision problems. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 345–352). Cambridge, MA: MIT Press.
- Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 99–134.
- Kemp, S. M. (1996). The language of animal learning theories: A radical behaviorist perspective. In J. Valsiner & H.-G. Voss (Eds.), *The structure of learning processes* (pp. 306–328). Norwood, NJ: Ablex.
- Kemp, S. M., & Eckerman, D. A. (1995). Direct analysis of contingencies using working models. *Revista Mexicana de Analisis de la Conducta*, 21, 27–46.
- Kintsch, W. (1970). Stochastic learning theory. In M. H. Marx (Ed.), *Learning: Theories* (pp. 195–234). New York: MacMillan.
- Lee, V. L. (1981). Terminological and conceptual revision in the experimental analysis of language development: Why. *Behaviorism*, 9, 25–53.
- Lee, V. L. (1983). Behavior as a constituent of conduct. *Behaviorism*, 11, 199–224.
- Lee, V. L. (1986). Act psychologies and the psychological nouns. *The Psychological Record*, 36, 167–177.
- Lee, V. L. (1988). *Beyond behaviorism*. Hillsdale, NJ: Erlbaum.
- Lee, V. L. (1996). Things done. *Revista Mexicana de Analisis de la Conducta*, 22, 137–158.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
- Mackintosh, N. J. (1983). *Conditioning and associative learning* (Oxford Psychology Series, No. 3). Oxford: Clarendon Press.
- Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, 27, 213–234.
- Meazzini, P., & Ricci, C. (1986). Molar vs. molecular units of behavior. In T. Thompson & M. D. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 19–43). Hillsdale, NJ: Erlbaum.
- Mechner, F. (1959). A notational system for the description of behavioral procedures. *Journal of the Experimental Analysis of Behavior*, 2, 133–150.
- Mechner, F. (1992). *The revealed operant: A way to study the characteristics of individual occurrences of operant responses*. Cambridge, MA: Cambridge Center for Behavioral Studies.
- Millenson, J. R. (1967). *Principles of behavior analysis*. New York: MacMillan.
- Monahan, G. E. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28, 1–16.
- Narendra, K. S., & Thathachar, M. A. L. (1989). *Learning automata: An introduction*. Englewood Cliffs, NJ: Prentice Hall.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- Norman, D. A. (1993). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science*, 17, 1–6.
- Norman, J. M. (1975). *Elementary dynamic programming*. London: Edward Arnold.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: Vol.*

2. *Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8, 164–212.
- SAS Institute. (1996). SAS/IML® [Computer software]. Cary, NC: Author.
- Schmajuk, N. A., Lamoureux, J. A., & Holland, P. C. (1998). Occasion setting: A neural network approach. *Psychological Review*, 105, 3–32.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922–932.
- Shull, R. L. (1979). The postreinforcement pause: Some implications for the correlational law of effect. In M. D. Zeiler & P. Harzem (Eds.), *Advances in analysis of behaviour: Vol. 1. Reinforcement and the organization of behaviour* (pp. 194–221). Chichester, England: Wiley.
- Shull, R. L. (1991). Mathematical description of operant behavior: An introduction. In I. H. Iversen & K. A. Lattal (Eds.), *Techniques in the behavioral and neural sciences: Vol. 6. Experimental analysis of behavior* (Part 2, pp. 243–282). Amsterdam: Elsevier.
- Shull, R. L. (1995). Interpreting cognitive phenomena: Review of Donahoe and Palmer's *Learning and Complex Behavior*. *Journal of the Experimental Analysis of Behavior*, 63, 347–358.
- Singh, S. P., Jaakola, T., & Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In W. W. Cohen & H. Hirsh (Eds.), *Proceedings of the 11th International Conference on Machine Learning* (pp. 284–292). San Francisco: Morgan Kaufmann.
- Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, 54, 193–216.
- Skinner, B. F. (1984a). Methods and theories in the experimental analysis of behavior. *Behavioral and Brain Sciences*, 7, 511–546.
- Skinner, B. F. (1984b). Reply to Harnad. *Behavioral and Brain Sciences*, 7, 721–724.
- Skinner, B. F. (1988a). Methods and theories in the experimental analysis of behavior. In A. C. Catania & S. Harnad (Eds.), *The selection of behavior: The operant behaviorism of B. F. Skinner: Comments and consequences* (pp. 77–105). Cambridge: Cambridge University Press.
- Skinner, B. F. (1988b). Reply to Harnad. In A. C. Catania & S. Harnad (Eds.), *The selection of behavior: The operant behaviorism of B. F. Skinner: Comments and consequences* (pp. 468–473). Cambridge: Cambridge University Press.
- Sondik, E. (1971). *The optimal control of partially observable Markov processes*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.
- Staddon, J. E. R. (1988). Quasi-dynamic choice models: Melioration and ratio invariance. *Journal of the Experimental Analysis of Behavior*, 49, 303–320.
- Staddon, J. E. R., & Zhang, Y. (1991). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. R. Staddon (Eds.), *Neural network models of conditioning and action: A volume in the quantitative analyses of behavior series* (pp. 279–293). Hillsdale, NJ: Erlbaum.
- Steinhauer, G. D. (1986). *Artificial behavior: Computer simulation of psychological processes*. Englewood Cliffs, NJ: Prentice Hall.
- Suchman, L. (1993). Response to Vera and Simon's situated action: A symbolic interpretation. *Cognitive Science*, 17, 71–75.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Vera, A. H., & Simon, H. A. (1993a). Situated action: A symbolic interpretation. *Cognitive Science*, 17, 7–48.
- Vera, A. H., & Simon, H. A. (1993b). Situated action: Reply to William Clancey. *Cognitive Science*, 17, 117–133.
- Weingarten, K., & Mechner, F. (1966). The contingency as an independent variable of social interaction. In T. Verhave (Ed.), *The experimental analysis of behavior: Selected readings* (pp. 447–459). New York: Appleton-Century-Crofts.
- Weiss, A. P. (1924). Behaviorism and behavior, I. *Psychological Review*, 31, 32–50.
- Weiss, A. P. (1925). *A theoretical basis of human behavior*. Columbus, OH: R. G. Adams.
- Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). New York: MacMillan.

Received March 18, 1999

Final acceptance January 29, 2001

APPENDIX A

Equations for the SZ Model

Stated formally, for situation i at time t :
 $R(t) = R_{ik}$, where $V_{ik}(t) \geq V_{ij}(t)$ for all deeds j (winner take all):
 with no reinforcement,

$$V_{ij}(t) = a_{ij}V_{ij}(t-1) + \epsilon(1 - a_{ij});$$

with reinforcement,

$$V_{ij}(t) = a_{ij}V_{ij}(t-1) + \epsilon(1 - a_{ij}) + b_{ij}V_{ij}(t-1),$$

$$0 < a_{ij} < 1; \quad a_{ij} + b_{ij} < 1;$$

where $V = \{v_{ij}\}$ is a matrix of values for each response in the response matrix, $R = \{r_{ij}\}$.

$A = \{a_{ij}\}$ is the matrix of adaptation parameters determining the duration of the effect of reinforcement. $B = \{b_{ij}\}$ is the matrix of arousal parameters determining the magni-

tude of the effect of reinforcement. ϵ is a random variable distributed uniformly over the $[0, 1]$ interval. $R(t)$ is the response made at time t (Staddon & Zhang, 1991).

APPENDIX B

Equations for the RE Model

Stated formally, for situation i at time t :

$$p_{ij}^{(t)} = \frac{v_{ij}^{(t)}}{\sum_j v_{ij}^{(t)}} \quad \text{for every deed, } j$$

$$a^{(t)} = a_{ij} \quad \text{with probability, } p_{ij}^{(t)}$$

$$W^{(t)} = \delta \cdot W^{(t-1)} \quad (\text{exponential decay})$$

$$w_{ij}^{(t)} = 1$$

$$V^{(t)} = V^{(t-1)} + c_{ij} \cdot \psi \cdot W^{(t)}$$

$$s^{(t+1)} = d_j^{(t)}(s_i),$$

where $V = \{v_{ij}\}$ is a matrix of values for each action in the action matrix, $A = \{a_{ij}\}$. $W = \{w_{ij}\}$ is the matrix of weights indicating the susceptibility of each action to subsequent reinforcement. δ is the decay rate, and ψ is the reinforcer potency. c_{ij} is 1.0 for a reinforcing action (e.g., feeding) and 0.0 for any other action. $a^{(t)}$ is the action taken at time t and $w_{ij}^{(t)}$ is the weight corresponding to that action.

APPENDIX C

Equations for the DMOD Model*

Notational conventions: Matrices are in uppercase italic, $G = \{g_{ij}\}$. There are n rows, one for each stimulus cue, and m columns, one for each deed. Vectors are lowercase boldface. The vector \mathbf{g}_j is the column vector for the j th column of G . $T()$ is transpose function. \circ is the Hadamard product operator. (Each element of the Hadamard product of two conformal matrices or arrays consists of the product of the corresponding elements of the factors.)

The Greek letters α , β , γ , and λ are reserved for parameters of the model and as such are constant across time. When indexed by the iteration cycle, t , it indicates that the choice of parameter varies over cycles as specified in the equations. All other variables are presumed to be indexed by the iteration cycle, t , unless marked with an asterisk, indicating the next cycle, $(t + 1)$. For example, \mathbf{s} is the column vector of stimulus cues present on cycle t , and \mathbf{s}^* is the column vector of stimulus cues present on cycle $t + 1$.

Parameters: γ is the spatial gradient param-

eter, ranging from 0 to 1, and is assumed constant for these simulations. α_i are the saliences, one for each stimulus cue. The vector $\boldsymbol{\alpha}$ is a column vector of the saliences. The β parameters are the learning and decay parameters for the various components for both primary and secondary reinforcement. There are 12 β parameters:

$\beta_{X,1}^+$	$\beta_{Y,1}^+$	$\beta_{Z,1}^+$	learning rates for primary reinforcement
$\beta_{X,1}^-$	$\beta_{Y,1}^-$	$\beta_{Z,1}^-$	decay rates for primary reinforcement
$\beta_{X,2}^+$	$\beta_{Y,2}^+$	$\beta_{Z,2}^+$	learning rates for secondary reinforcement
$\beta_{X,2}^-$	$\beta_{Y,2}^-$	$\beta_{Z,2}^-$	decay rates for secondary reinforcement

although, for most simulations, all learning rates are equal and all decay rates are equal. λ is the value for primary reinforcement, set to 1 in these simulations. $\lambda(t) = \lambda$ when reinforcement occurs and 0 otherwise.

The action, $a(t)$, on each cycle is determined based upon the matrix of values, $V = \{v_{ij}\}$ on cycle t . At every point, V is composed

of three additive components: approach, X , avoidance, Y , and counterconditioning, Z . These components are altered by both primary and secondary reinforcement on each cycle.

For situation k at time t :

$$V = X + \gamma Y + Z,$$

$\mathbf{q} = T(\mathbf{s})V$, where $s_i = 1$ if cue i is present and 0 otherwise.

$$\mathbf{u} = (|\mathbf{q}| + \mathbf{q})/2$$

(set all negative values to 0)

$$a^{(t)} = a_{ij}$$

with probability

$$p_j = \frac{u_j}{\sum_j u_j}.$$

Primary reinforcement, given deed $j = a_{ij}$, for situation i at time t :

$$V = (X + \Delta X) + \gamma(Y + \Delta Y) + (Z + \Delta Z),$$

where $\Delta X, \Delta Y, \Delta Z$ are 0 for all columns $\neq j$ and

$$\Delta \mathbf{x}_{\cdot j} = (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{X,1}^{(t)}(d_X)$$

where

$$d_X = \lambda(t) - T(\mathbf{s})\mathbf{x}_{\cdot j} \quad \text{and}$$

$$\beta_{X,1}^{(t)} = \begin{cases} \beta_{X,1}^+ & \text{if } d_X \geq 0 \\ \beta_{X,1}^- & \text{otherwise.} \end{cases}$$

$$\Delta \mathbf{y}_{\cdot j} = \begin{cases} (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Y,1}^{(t)}(d_Y) & \text{if } d_X < 0 \\ (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Y,1}^{(t)}(0 - T(\mathbf{s})\mathbf{y}_{\cdot j}) & \text{otherwise,} \end{cases}$$

where

$$d_Y = 2d_X - T(\mathbf{s})\mathbf{y}_{\cdot j} \quad \text{and}$$

$$\beta_{Y,1}^{(t)} = \begin{cases} \beta_{Y,1}^+ & \text{if } d_X, d_Y < 0 \\ \beta_{Y,1}^- & \text{otherwise.} \end{cases}$$

$$\Delta \mathbf{z}_{\cdot j} = \begin{cases} (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Z,1}^{(t)}(d_Z) & \text{if } d_X \geq 0 \\ (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Z,1}^{(t)}(0 - T(\mathbf{s})\mathbf{z}_{\cdot j}) & \text{otherwise,} \end{cases}$$

where

$$d_Z = 0 - T(\mathbf{s})\mathbf{y}_{\cdot j} - T(\mathbf{s})\mathbf{z}_{\cdot j} \quad \text{and}$$

$$\beta_{Z,1}^{(t)} = \begin{cases} \beta_{Z,1}^+ & \text{if } d_X \geq 0 \quad \text{and} \quad d_Y > 0 \\ \beta_{Z,1}^- & \text{otherwise.} \end{cases}$$

Secondary reinforcement:

$$\mathbf{e} = T(\mathbf{s}^*)V$$

where $s_i^* = 1$ if cue i is present at time $(t + 1)$ and 0 otherwise.

$$\mathbf{c} = (|\mathbf{e}| + \mathbf{e})/2$$

(set all negative values to zero)

$$\mathbf{w} = \frac{\mathbf{c}}{\sum_j c_{ij}}.$$

Given deed $j = a_{ij}$, for situation i at time t :

$$V = (X + \Delta X) + \gamma(Y + \Delta Y) + (Z + \Delta Z),$$

where $\Delta X, \Delta Y, \Delta Z$ are 0 for all columns $\neq j$ and

$$\Delta \mathbf{x}_{\cdot j} = (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{X,2}^{(t)}(\mathbf{w}T(X)\mathbf{s}^* - T(\mathbf{s})\mathbf{x}_{\cdot j})$$

where

$$\beta_{X,2}^{(t)} = \begin{cases} \beta_{X,2}^+ & \text{if } \mathbf{w}T(X)\mathbf{s}^* \geq T(\mathbf{s})\mathbf{x}_{\cdot j} \\ \beta_{X,2}^- & \text{otherwise.} \end{cases}$$

$$\Delta \mathbf{y}_{\cdot j} = (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Y,2}^{(t)}[\mathbf{w}T(Y)\mathbf{s}^* - T(\mathbf{s})\mathbf{y}_{\cdot j}]$$

where

$$\beta_{Y,2}^{(t)} = \begin{cases} \beta_{Y,2}^+ & \text{if } \mathbf{w}T(Y)\mathbf{s}^* \leq T(\mathbf{s})\mathbf{y}_{\cdot j} \\ \beta_{Y,2}^- & \text{otherwise.} \end{cases}$$

$$\Delta \mathbf{z}_{\cdot j} = (\mathbf{s} \circ \boldsymbol{\alpha}) \beta_{Z,2}^{(t)}[\mathbf{w}T(Z)\mathbf{s}^* - T(\mathbf{s})\mathbf{z}_{\cdot j}]$$

where

$$\beta_{Z,2}^{(t)} = \begin{cases} \beta_{Z,2}^+ & \text{if } \mathbf{w}T(Z)\mathbf{s}^* \geq T(\mathbf{s})\mathbf{z}_{\cdot j} \\ \beta_{Z,2}^- & \text{otherwise.} \end{cases}$$