

The Process of Recurrent Choice

D. G. S. Davis, J. E. R. Staddon, A. Machado, and R. G. Palmer

Recurrent choice has been studied for many years. A static law, *matching*, has been established, but there is no consensus on the underlying dynamic process. The authors distinguish between dynamic models in which the model state is identified with directly measurable behavioral properties (performance models) and models in which the relation between behavior and state is indirect (state models). Most popular dynamic choice models are local, performance models. The authors show that behavior in different types of discrimination-reversal experiments and in extinction is not explained by 2 versions of a popular local model and that the nonlocal *cumulative-effects model* is consistent with matching and that it can duplicate the major properties of recurrent choice in a set of discrimination-reversal experiments. The model can also duplicate results from several other experiments on extinction after complex discrimination training.

Choice has been a major theme in operant research for the past 30 years. Pigeons, rats, and human beings have been studied on a variety of procedures in which simple repetitive responses, such as key pecking or lever pressing, are intermittently rewarded (reinforced) according to various schedules. The allocation of behavior among choice alternatives, typically left or right keys or levers, has been measured as a function of the type of schedule, schedule parameters, and the obtained rate of reinforcement (see Williams, 1988, and various chapters in Honig & Staddon, 1977, for reviews).

For most of this 30-year period, theoretical emphasis has been on static, reversible, molar equilibrium principles, chief among which is the *matching law* (Davison & McCarthy, 1988; Herrnstein, 1961). Matching is an empirical relation demonstrated most clearly in choice between two or more variable-interval (VI) reinforcement schedules (i.e., schedules in which the first response after a variable time since the preceding reinforcement is reinforced). The finding here is that under appropriate steady-state conditions, the ratio of response rates, x/y , is approximately equal to the ratio of obtained reinforcement rates, $R(x)/R(y)$, so that $x/(x + y) = R(x)/[R(x) + R(y)]$. (It is worth remembering for later discussions that although matching is usually expressed as equality of response and reinforcement proportions, it can just as well be expressed as equality of reinforcement probabilities: $R(x)/x = R(y)/y$.)

The matching law is reversible in the sense that matching is usually a stable end state under given conditions of reinforcement, largely independent of the organism's prior history. It is molar because the rates x , y , $R(x)$, and $R(y)$ are measured over an extended period, typically several hours (e.g., the average of the last five experimental sessions under a given set of schedules). In addition, it is an equilibrium principle (rather than a causal law) because it relates two quantities, response and reinforcement rates, that are mutually dependent.

The static matching relation is compatible with many dynamic choice processes (cf. Hinson & Staddon, 1983). Nevertheless, an understanding of the specific process or processes that underlie choice would have several advantages. It would help us to explain not only matching but perhaps also systematic deviations from matching. It would allow us to say something about choice on a moment-by-moment basis—ideally in real time, but if not, at least on a response-by-response basis. It would open the way to understanding behavior that depends on remote past history, that is, on experimental conditions preceding the current one. Also, it might tell us something about the transfer properties of a particular training history, such as its effects on behavior in extinction or on the learning of some new task.

In this article, we attempt to identify essential properties of the process of recurrent choice. The article is in five major sections that constitute a single argument. In Part I, we discuss types of dynamic models. We introduce the distinction between *performance* models (i.e., models where the relevant variables can be directly measured) and *state* models (i.e., models where the relevant variables must be inferred from historical information). We also make a distinction between models that are *local* and *nonlocal*. This discussion is necessary because we want to argue that the properties of choice imply a nonlocal, state model, whereas the major dynamic models for recurrent choice are local, performance models.

In Part II, we discuss the dominant type of local choice model, the "leaky integrator." We remind the reader that a performance version of the integrator provides a very good fit to extensive, individual-subject data on daily serial-reversal learn-

D. G. S. Davis, J. E. R. Staddon, and A. Machado, Department of Psychology: Experimental, Duke University; R. G. Palmer, Department of Physics, Duke University.

Research support was provided by grants to Duke University from the National Science Foundation and the National Institute of Mental Health (J. E. R. Staddon, principal investigator). We thank Jennifer Higa, Nancy Innis, Arata Kubota, Ben Williams, and Clive Wynne for comments on earlier versions of this article and all the members of the Learning and Adaptive Behavior group for many discussions of these issues. J. E. R. Staddon also thanks the Alexander von Humboldt-Stiftung for support.

Correspondence concerning this article should be addressed to J. E. R. Staddon, Department of Psychology: Experimental, Duke University, P.O. Box 90086, Durham, North Carolina 27706-0086.

ing. We show that the performance version nevertheless fails to predict three other well-established properties: regression, improvement across a reversal series, and systematic differences between reversal every day versus every 2 or every 4 days. We go on to compare this performance integrator with a winner-take-all state version, which can handle some aspects of regression but shares the other defects of the performance version.

The comparison between the performance and state versions of the integrator allows us to argue (a) that the winner-take-all response rule is critical to the regression property and should therefore form part of any improved model and (b) that the failure of both integrator models to account for the other properties of serial-reversal learning implies that the underlying process is nonlocal.

In Part III, we introduce a simple new model, the *cumulative-effects* (CE) model, which is nonlocal and includes a winner-take-all response rule. We show in that section that the CE model overcomes the main defects revealed by our discussion of the two integrator models in Part II.

In Part IV, we discuss the strengths and weaknesses of the CE model in relation to other choice data. Part V is a brief conclusion section.

Our general objective is to explain the effect of complex histories on the choice behavior of individual subjects. Our strategy is to compare simple computable models, hoping to learn from the successes and failures of particular models what model properties are critical to a full understanding of choice behavior. By putting the critical properties together in various combinations, we hope to arrive at increasingly comprehensive models for choice.

The effort to explain complex histories is in one way more ambitious than the usual task attempted by theories of choice: explaining reversible steady-state relationships or, at most, the effects of single transitions, such as acquisition or extinction functions, or the transition between a discrimination and its reversal. The trade-off is that we will be content at this stage to make correct qualitative predictions of an extensive data set, rather than exact quantitative predictions of a more restricted set.

1. Performance and State Models

The attempt to understand behavioral phenomena through theory involves the simultaneous exploration of two domains: the domain of experimental results and the domain of possible theoretical systems. A successful theory represents the correct identification of isomorphic regions in these two domains: a theoretical system that closely matches a set of experimental results. This search of two infinite domains is impossible without classification. Experimental results are typically classified procedurally—classical versus instrumental conditioning, choice versus nonchoice procedures, and so on. There is less consensus on the proper classification of theoretical systems. One way to classify dynamic choice models is as follows.

We consider only deterministic, computable models, that is, models in which the state in the next instant is a well-defined function of the current state and input. (We are not concerned with informal or purely verbal models or with stochastic mod-

els.) Any computable model for recurrent choice can be reduced to the following discrete form:

$$\mathbf{Y}(t+1) = \mathbf{f}[\mathbf{Y}(t), \mathbf{R}(t)], \quad (1)$$

where $\mathbf{Y}(t)$ is a vector representing the state of the model and $\mathbf{R}(t)$ is a vector representing the reinforcement conditions at time (or iteration) t . In other words, a computable choice model gives the state of the model in the next iteration as a function of the state and reinforcement conditions in the preceding iteration.

Equation 1 is the *model definition*. A second function, the *response rule*, maps the state of the model onto behavior:

$$\mathbf{B}(t) = \mathbf{g}[\mathbf{Y}(t)], \quad (2)$$

where \mathbf{B} is some measurable behavioral property such as choice proportion or response occurrence.

The critical issue for the distinction between performance and state models is whether vector \mathbf{Y} can be eliminated from these two equations and be replaced by a function of \mathbf{B} alone. If \mathbf{Y} can be eliminated, the model is a performance model. The simplest possibility is that function \mathbf{g} in Equation 2 is simply a one-one mapping, so that the state of the model is uniquely defined by some measurable property of behavior. For example, \mathbf{Y} could be defined as response rate or choice proportion measured over some specified window. Equation 2 is obviously redundant for performance models because \mathbf{Y} in Equation 1 can be rewritten in terms of directly measurable quantities.

Function \mathbf{g} can also be a many-one mapping; that is, a given model state defines a unique behavior, but a given behavior may be compatible with more than one model state. Unless the variables can be redefined in such a way as to allow \mathbf{Y} to be eliminated, such models are state models (see Appendix A for a discussion of the conditions under which this redefinition is possible). For example, any model that says that a response occurs whenever some variable exceeds a threshold is a state model because \mathbf{Y} , the state of the model, cannot be uniquely identified from the fact that a response did, or did not, occur on a particular iteration.

Performance models correspond to what has been termed the *independence-of-path* assumption, that is, the idea that future behavior depends only on the current behavior and input. What our classification adds is a distinction between the properties of the model (which always shows independence of path in state space) and the relation between the model state and observable quantities, which depends on the *response rule* (see Appendix A).

Scale

We can also classify models that are based on the scale of t in Equations 1 and 2 (this classification is similar to one proposed by Sutton, 1984). If each iteration is "clocked" in real time, brief instant by brief instant, then function \mathbf{g} takes the form of a decision among available responses (including "no response"): At each instant, t , just one of the available responses can occur. This makes the model a *real-time* model, the finest scale.

The index variable, t , can also iterate asynchronously, response by response. In this case also, function \mathbf{g} is a decision among the available responses, a binary decision in the two-

choice case. The model simply specifies, response by response, whether the choice will be left or right. The scale of this type of model might be termed *response-level*. Obviously, a given real-time model implies a response-level model but not conversely.

Finally, t can be iterated over some larger period of time or responses such as days or experimental sessions. In this case, function g represents a statistical aggregate, such as a mean rate or proportion. Models of this type are also termed *molar*. A molar model might predict the proportion of right choices today given the proportion yesterday and the reinforcement conditions today, for example. Other molar models predict choice probability or response rate over a more or less well-defined time or response window. Again, there is an inclusive upward relation: A response-level model implies a molar model but not conversely.

Examples

Herrnstein and Vaughan's (1980) *melioration* is one example of a molar, performance model:

Subjects compare local rates of reinforcement from concurrently available alternatives and shift in a relatively continuous manner towards the higher one. . . . the process itself appears to be psychologically simple, requiring the subject to detect nothing more than signed differences in local reinforcement rate. (Herrnstein & Vaughan, 1980, p. 164)

The state of this model is identified with local response and reinforcement rates that, once the word "local" is defined, correspond to directly measurable behavioral properties. Other performance models are the *kinetic model* of Myerson and Miezin (1980; states = rates) and the *ratio invariance* model of Horner and Staddon (1987; states = probabilities).¹

Harley's Relative-Payoff Sum rule (Harley, 1981) provides an example of a state model. The model definition is $Y(t+1) = aY(t) + (1-a)Y(0) + R(t)$, $0 < a < 1$, where Y is a vector of response strengths and R represents the reinforcement conditions. The response rule is $B_i(t) = Y_i(t)/[\sum_j Y_j(t)]$ where $B_i(t)$ is the probability of response i and Y_i is the i th component of vector Y . Given the form of the functions f and g , $B(t+1)$ cannot be expressed as a function of $B(t)$ and $R(t)$ alone.

State Identification

The lack of isomorphism between the state of a state model and any single behavioral property does not mean, of course, that the state of the model can never be identified. For many state models, a particular sequence or sequences of experimental inputs (a particular history or set of histories) may be sufficient to bring the model to a known state from which all future states will then be predictable. However, this may not be possible for every state model.

It will also often be possible to combine information from behavioral observations made at different times to identify the current state of the model. This is just another way of saying that, from the point of view of an external observer, the state of a state model is nothing more than a set of equivalent (behavioral) histories, equivalent in the sense that the future behavior of the system is the same following any of the histories in the set

(cf. Minsky, 1967, for a clear exposition of this fundamental property; see also Staddon, 1973).

Locality

The term local is often used to describe choice models (cf. the aforementioned melioration example), but the term is not well defined. A simple definition that seems to make sense is as follows. Consider the question, "How much historical information is necessary to predict the future behavior of the model (i.e., to define its state)?" Clearly, for all performance models only the current performance and reinforcement conditions are necessary, because they define the model state. However, state models can differ in the amount of historical information necessary to define the current state. When the information required includes only events in the recent past, we define the model as local, but when information about the remote past is necessary, we define the model as nonlocal. Thus, all performance models are local, but only some state models are local. We give examples in Part II.

Note that locality is a property of a model, not of behavior. It is not possible therefore to identify particular behavior as locally or nonlocally determined in the absence of any model for the behavior.

Most of the process models that have been proposed for recurrent choice are local in the present sense. All six models studied in a comprehensive review by Dow and Lea (1987), for example, seem to be local in this sense. In this article, we contrast three choice models: two local models, performance and state versions of a leaky integrator, and a nonlocal model, which we call the *cumulative-effects* (CE) model. On the basis of the failures of the two local models and the relative success of the nonlocal model, we argue that the data we discuss seem to require a nonlocal process.

II. Integrator Model

The "leaky integrator," first introduced into psychology by Bush and Mosteller (1955), is by far the most popular process that has been used to explain reinforcement learning. We discuss two versions of the integrator that in various forms have been widely used to explain recurrent choice.

The integrator works in discrete time as follows. Suppose that the strength of a given response in the next instant of time, $V(t+1)$, depends only on its current strength, $V(t)$, and whether or not a stimulus occurred at time t , that is,

$$V(t+1) = aV(t) + (1-a)X(t), \quad 0 < a < 1, \quad (3)$$

where $X(t)$ is stimulus magnitude at time (or iteration) t and a , the time constant,² is a parameter that represents the persistence of effects. This process has also been termed the *linear*

¹ This is a slight oversimplification, because there are often severe technical difficulties in going back from empirical choice proportions to the setting on the model's probability generator. These problems arise with any stochastic model. We do not deal with stochastic models in this article.

² Strictly speaking, the time constant for the corresponding exponential is $1/(1-a)$, but we refer to a for simplicity.

model, an exponentially weighted moving average (Killeen, 1981), or the common model (Lea & Dow, 1984). The state of this process, in our terminology, is just the value of V .

The response, $V(t)$, of the integrator to a brief, "spike" input and a longer, "square-wave" input, $X(t)$, is shown in Figure 1.

As Figure 1 shows, the effects on strength, $V(t)$, of any change in the value of X decline to a negligible level at a rate determined by the time constant, a . Thus, information about past events is not retained indefinitely, and the model returns as closely as we like to the zero state with sufficient lapse of time. This is a local model in our classification.

Reversal Learning: Two Versions of the Integrator

Simple as it is, the ability of a purely performance-model integrator to describe some molar choice data is quite impressive. An example is shown in Figure 2. The figure shows data from a two-armed-bandit (concurrent random ratio) experiment (Davis & Staddon, 1990) in which hungry pigeons chose between two response keys, each paid off with food reinforcement according to probabilistic schedules. The pigeons were rewarded each day, with probability $1/8$, for pecking one of two keys. The "hot" key varied from day to day in an irregular fashion (left = L, right = R, etc. at the top), and some days only forced alternation (F) was rewarded. On a few days, neither response was rewarded (extinction = E) or the animal was not run (O).

The triangles are data points: average daily choice proportion $S = R/(R + L)$ for a single pigeon (in what follows, we use S to denote empirical choice proportions and s to denote theoretical predictions). The solid line shows the fit of the integrator,

$$s(t + 1) = as(t) + (1 - a)X(t + 1), \quad (4)$$

where s = predicted choice proportion and X is the asymptote each day. X is set each day according to the known effects of maintaining each daily condition indefinitely under these conditions: $X = 1$ (R), 0 (L), or 0.5 (E, O, and F). Predicted proportion, s , tracks the daily changes in actual proportion, S , well.

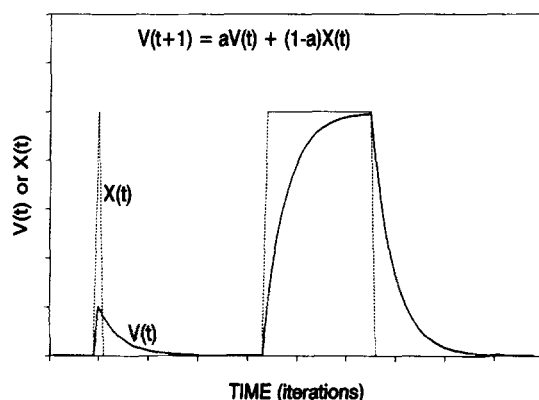


Figure 1. The effect of a "spike" stimulus and a "square-wave" stimulus (dashed, light line, $X(t)$) on the state of a leaky integrator ($V(t)$, heavy line, Equation 3 in the text). (t = time, or iteration; a = time constant.)

This good fit by a local model suggests that the choice process under these conditions is essentially local. Despite the good fit of the integrator in this experiment, the essential properties of choice are probably not local, however.

Molar integrator model (INT-M). The good fit of the integrator here conceals some problems. The most obvious is that the response here is a proportion, s , the proportion of pecks on the right key. There is no evidence that real organisms learn proportions, but there is much evidence that they learn particular responses.³ Thus, to be realistic, the model needs to be modified so that there are at least two integrators, for left and right responses, with some rule for combining them. The model then becomes

$$V_L(t + 1) = aV_L(t) + (1 - a)X_L(t + 1) \quad \text{and} \quad (5a)$$

$$V_R(t + 1) = aV_R(t) + (1 - a)X_R(t + 1), \quad (5b)$$

and the response rule is

$$s = V_R/(V_R + V_L), \quad (6)$$

where X_L and X_R are asymptotes set equal to 1 when that response is reinforced and 0 if it is not, and s is the predicted proportion of right choices. Following Dow and Lea (1987), we term Equation 6 a *matching* response rule. If we add the constraint that $V_L(0) + V_R(0) = 1$, the net result in this situation is the same as the one-integrator model, but the two-integrator version makes more conceptual sense and makes predictions for the case where neither response is reinforced.

Response-by-response integrator model (INT-R). The response-by-response version is as follows. For each of the two choice alternatives, response strength V is computed as in Equation 3: $V(t + 1) = aV(t) + (1 - a)X(t)$, but the computations take place response by response, not session by session. If a left response (say) is made, then $V_L(t)$ is updated according to Equation 3, with the asymptote determined by whether the response was reinforced: If reinforced, $X_L(t)$ is set equal to the magnitude of reinforcement on that occasion (one, in all the simulations here); if unreinforced, $X_L(t)$ is set equal to zero. $V_R(t)$ is not changed because a right response did not occur.

The functional rationale for the assumption that the value of a *silent* (i.e., nonoccurring) response remains constant is that in the absence of any time-telling process the organism has information about the value of a choice only when the response occurs. (This assumption may need to be abandoned in a real-time model.)

The response rule for INT-R is winner take all (also termed a *maximizing rule*); that is, the response with the highest V value is the response that occurs on that iteration. With this rule, INT-R corresponds to what Dow and Lea (1987) called a "dynamic meliorator." INT-M is a performance model in our classification; INT-R is a state model (because the highest-wins rule

³ For example, if animals can learn choice proportions (rather than simply learning to peck left or right), then they should adapt well to frequency-dependent schedules in which particular choice proportions are preferentially rewarded. However, in fact, they routinely fail to maximize reinforcement rate on such procedures (e.g., Horner & Staddon, 1987; Machado, 1992; Mazur, 1981; Staddon & Hinson, 1983).

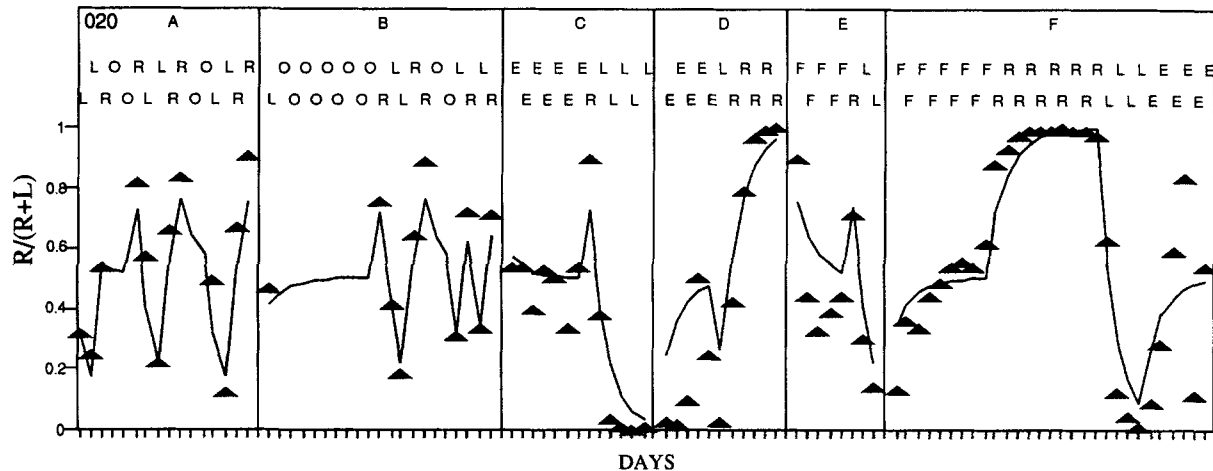


Figure 2. The entire course (Phases A–F) of a discrimination-reversal experiment with a single pigeon. (The solid triangles show the proportion of total responses to the right key each day when reinforcement conditions [shown by the staggered letters at the top] are changed frequently. The conditions were L [pecks on the left key paid off with probability $1/8$], R [pecks on the right key paid off with probability $1/8$], E [extinction], F [alternations only, $L \rightarrow R$ or $R \rightarrow L$, paid off with probability 1] or O [animal not run that day]. The solid line is the best-fitting prediction of Equation 4 in the text [$\alpha = 0.55$; the correlation between predicted and obtained values is 0.926]. [Data from Davis & Staddon, 1990, which should be consulted for experimental details.]

means that the state of the silent response is not directly revealed in behavior).

Notice that because of the assumption that V values are only updated when a response occurs, the response-level model, INT-R, is only conditionally local. The model states cannot in fact be recovered under usual conditions because of a sort of “ratchet” effect related to the winner-take-all response rule, so that the current state of the model always reflects the entire history of the experiment (see Appendix B). Nevertheless, successive states (V values) become increasingly similar under some conditions, so that the model is approximately local in our terms.

We now explore the predictions of INT-M and INT-R for three experimental choice situations involving discrimination reversal: (a) a single reversal followed by extinction—conditions that can lead to regression, a form of spontaneous recovery; (b) successive daily discrimination reversal—conditions that normally lead to improvement in the rate of acquisition each day (i.e., so-called discrimination-reversal set); and (c) reversal after blocks of days—conditions that lead to slower reversal performance after longer training blocks. We also discuss the predictions of each model for concurrent variable-interval schedules—conditions that lead to matching of response and reinforcer ratios.

Regression

Models. Suppose that instead of random daily alternation between left and right reinforcement, as in Figure 2, we reinforce in two long blocks, first right-only reinforced for several sessions, then left-only reinforced, and finally extinction (neither reinforced): R–L–E. What do INT-M and INT-R predict? The prediction of INT-M is shown in Figure 3. The figure

shows V values for left (+), right (open squares), and the proportion of right choices (solid line). Both V values were set equal to 0.5 at the beginning of training. The extinction prediction is straightforward: The response proportion at the end of the last training condition is maintained indefinitely in extinction. (Response rates of course should decline; but the output of INT-M is determined by the ratio of V values, which, for two exponentials with the same time constant, will always remain constant.)

The behavior of INT-R is more complicated, and its predictions in this situation are quite different. Figure 4 shows the first 100 iterations of INT-R as it learns to choose the right

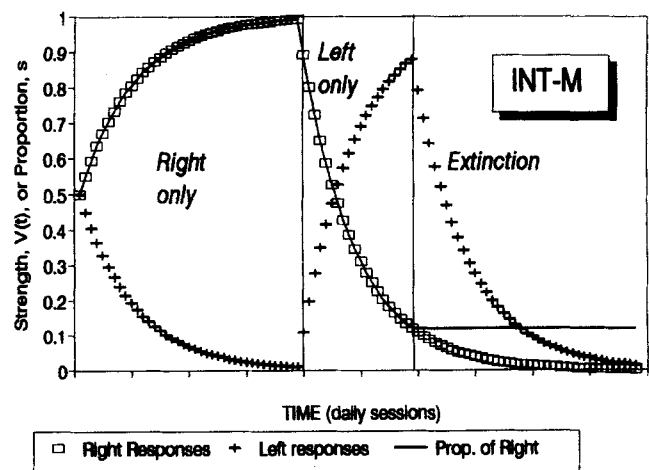


Figure 3. Molar integrator model (INT-M) predictions for discrimination reversal (several sessions of right only followed by several left only and then extinction: R–L–E. Prop. = proportion.)

response at the beginning of the right-left-extinction (R-L-E) sequence; right is here reinforced with probability 0.1. The figure shows cumulative right responses plotted against iterations, with the V value for the right on the right-hand y -axis. Because either a right or a left response occurs on each iteration in this response-by-response model, a horizontal slope indicates a left response and a 45° slope indicates a right response. The initial V values for both responses were 0.5. The figure shows how each reinforcement ("o" symbols) transiently boosts the V value for the reinforced response, which then occurs exclusively for the next few iterations; otherwise, the two responses alternate.

Because this model predicts response by response, its behavior can be seen most easily in the form of a trajectory in which cumulative left responses are plotted against cumulative right responses. The complete R-L-E sequence is shown in this form in Figure 5. As with INT-M, preference shifts toward the reinforced alternative, first right, then left. However, in extinction the current preference is only maintained briefly. After this brief period, there is an abrupt transition to approximate alternation between the two choices. This result is independent of the value of a and of the durations or reinforcement probabilities associated with phases R and L. It is analyzed formally in Appendix B. This reversion to a less extreme preference in extinction is *regression*, an apparently spontaneous recovery of an earlier preference.⁴

Data. What are the empirical results in situations like this? Results of several sorts are in the literature. Nevin, Tota, Torquato, and Shull (1990), for example, described inconsistent results in their experiment and earlier ones: Sometimes preference in extinction shifts towards the alternative most recently associated with the higher reinforcement rate, sometimes preference approaches indifference, and sometimes preference remains constant. Myerson and Hale (1988) showed that in extinction following concurrent VI-90-VI-180-s training, pigeons' preference remained approximately constant, as predicted by the INT-M, rather than drifting towards indifference.

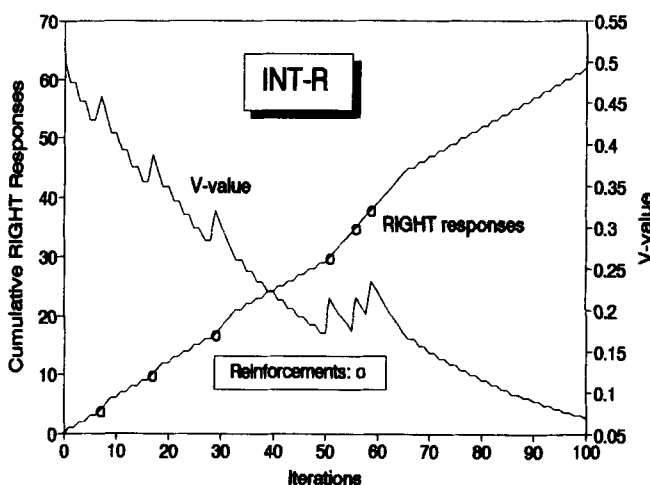


Figure 4. Response-by-response integrator model (INT-R) when right choices are rewarded with probability 0.1 (100 iterations). (Solid line with o symbols is cumulative right choices. Declining solid line is the V value for right. $V_L(0) = V_R(0) = 0.5$; time constant, $a = 0.95$.)

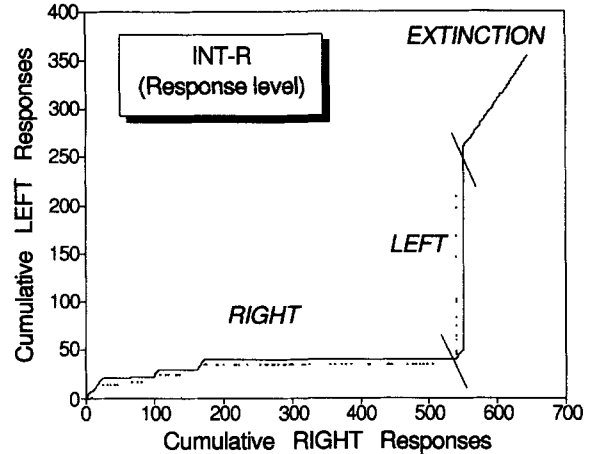


Figure 5. Response-by-response integrator model (INT-R) behavior during and following discrimination reversal: right (600 iterations), left (200 iterations), extinction (400 iterations): R-L-E. (Dots below [right] or above [left] the line indicate reinforcements. After a brief perseveration on the left at the beginning of extinction, the model alternates between left and right. This is an example of regression to an earlier pattern.)

ence, as we assumed in fitting Equation 4 in Figure 2 and as predicted by the INT-R in Figure 5. This constant preference was appropriate to the overall experience of the animals, however. Myerson and Hale described their animals' histories thus: "Prior to this [i.e., the concurrent (CONC) VI-90-VI-180-s] condition, all four birds had spent one month on a CONC VI VR (variable ratio) reinforcement schedule that provided comparable relative reinforcement rates" (p. 296). We cannot tell from these data, therefore, whether preference in extinction would be unchanging if the preextinction reinforcement condition produced a preference different from earlier conditions. When it does, a more common result than constancy of preference seems to be regression, however. The last panel in Figure 2 shows an example in the last E period: This animal began extinction with an almost exclusive left preference, but over sessions, his preference shifted through indifference to a right preference by the fourth session of extinction. A similar, but smaller, regression effect is apparent in the extinction in Panel D.

INT-R can only predict indifference as the terminal preference (i.e., $s \approx 0.5$) in extinction. INT-M must predict persistence of the preextinction preference. Thus, neither can accommodate the full range of empirical results, which show that preference in extinction depends to some extent on the animal's experience before the preceding reinforcement condition. INT-M, a local, performance model, is completely insensitive to experience earlier than the reinforcement condition that pre-

⁴ There is another form of spontaneous recovery in which a response that has been extinguished recovers solely as a function of lapse of time. The models we discuss in this article are computed response by response or session by session and have no explicit representation of time. Hence, they cannot deal with time-based spontaneous recovery.

cedes extinction. INT-R is only weakly sensitive to earlier experience: The longer (up to a point) the preceding reinforcement condition, the longer the model persists in its last preference before it switches instantaneously to indifference. The "regression" shown by INT-R is not really an expression of prior training, it is just the invariable response of the model to continued extinction.

Despite the problems with INT-R, the difference between INT-R and INT-M is instructive. The reason that INT-R allows for a shift in preference in extinction is the winner-take-all response rule plus the assumption that the strength of an unexpressed, "silent" activity remains constant. These two properties embody something like Clark Hull's (1934) "habit-family hierarchy." Available responses in a given situation have differential strengths, depending on their individual reinforcement histories. The dominant response is weakened in extinction, which allows the next in strength to show itself. We therefore preserved these two assumptions in the cumulative effects model (to be discussed later), which overcomes many of the limitations of INT-R.

Successive Daily Reversal (SDR)

Models. The molar model, INT-M, fails to predict improvement across successive daily discrimination reversals. If the initial value for V_L and V_R is different from the steady-state values, the model shows a shift in mean s value across days; however, it never shows progressive improvement in discrimination performance: the excursion in s value from day to day does not increase with experience as it should if discrimination performance is to improve across successive reversals.

The predictions of INT-R are again more complex, but we have found no conditions under which it gives a reliable improvement in discrimination performance across reversals. The first reversal (second discrimination day) is sometimes better than the first, but there is never systematic improvement after that point.

Data. Figure 6 shows typical data from daily discrimination reversal in the two-armed-bandit situation. The top two graphs are from pigeons with extensive experience in a choice situation, whereas the bottom two graphs are from experimentally naive birds. The solid lines show discrimination performance (percentage correct) each day. The dashed lines show the a parameter in Equation 4, computed from the data as shown in the figure legend.⁵ This value should not vary systematically if the animals learn at the same rate each day. It does vary. After two or three reversals, both naive and experienced birds show progressive improvement in performance, and the computed a value decreases, indicating faster discrimination learning each day. The naive animals show a drop in performance on the 2nd and 3rd days (first and second discrimination reversals), however. This pattern has also been found in earlier daily-reversal-learning studies (e.g., Staddon & Frank, 1974). Neither of the INT models can account either for the progressive improvement in all the birds or for the reliable difference between experienced and naive animals during early reversals.

Reversal in Blocks

Data and models. Pigeons reverse faster after exposure to daily reversals than to reversals in blocks of 2 or 4 days. These

data are shown for a single pigeon in Figure 7. The figure shows three effects shown by all 4 pigeons in this experiment:

1. Improvement in reversal performance across SDRs. This is shown both by daily improvement in percentage correct responses (solid line) and by the increased learning rate (reduced value of a , computed from Equation 4, dashed line) across the series of daily reversals (Single Alternation) in the left-hand panel and on the far right of the right-hand panel.

2. Change in learning rate (parameter a in Equation 4) as a function of frequency of reversal: a is lowest in daily reversal and highest when reversal is every 4 days.

3. Change in learning rate within and between blocks in the 2- and 4-day-reversal conditions: the computed a value increases within each block and then decreases between blocks.

These results are obviously incompatible with INT-M, because the a parameter computed from the data is not constant but varies systematically within and between conditions. We have already seen that INT-R is unable to account for progressive improvement in a series of daily reversals; it also cannot account for the changes in learning rate within blocks when reversal occurs at 2- or 4-day intervals.

Matching

None of the models we discuss in this article incorporates time. Nevertheless, we can make predictions about concurrent VI schedules by assuming that a left or right response occurs in each t s of real time. Here, and in simulations with the CE model, we set $t = 1$ s.

The molar model, INT-M, makes no prediction for concurrent VI-VI choice because the expected asymptote, $X(t+1)$, for each choice must be known in advance. The predictions for INT-R are surprisingly complex. Under most conditions, it predicts undermatching, that is, choice ratios, x/y , that are less extreme than reinforcement ratios, $R(x)/R(y)$. INT-R behaves strangely in one respect: With increasing experience, the length of runs (successive response on one key) grows without limit, an irreversible pattern quite different from any data. The reason for this curious behavior (which is not hinted at in other accounts of this model) is explained in Appendix B.

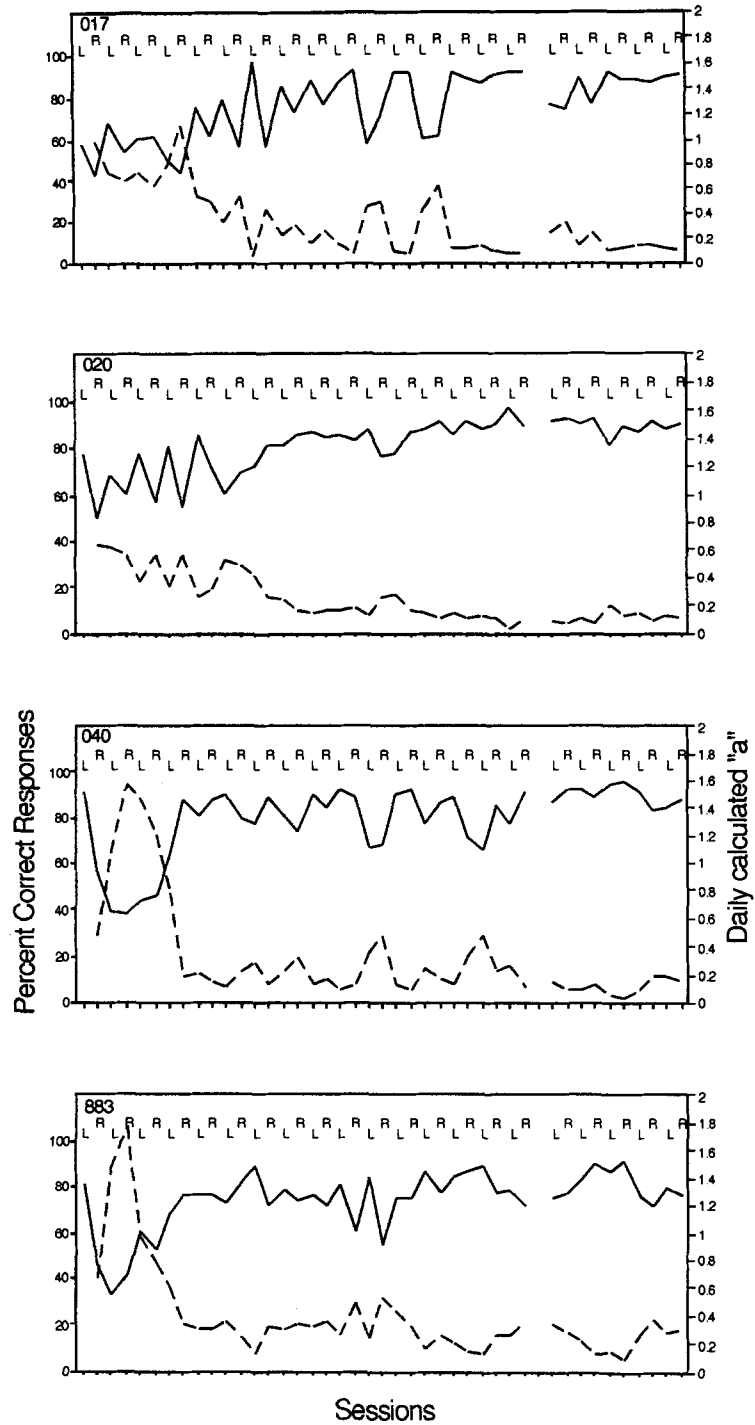
In summary, only the response-by-response, winner-take-all version of the integrator, INT-R, can predict approximate matching and regression effects in extinction, but neither model can account for improvement in performance across discrimination reversals or for differences in performance when reversals occur daily or in blocks. In the next section, we discuss a simple nonlocal model that begins to remedy some of these deficiencies.

III. Cumulative-Effects (CE) Model

The regression effect in extinction, which was predicted to some extent by INT-R, seems to depend on the WTA response

⁵ Recall that INT-M provided a very good description for the behavior of experienced animals on successive discrimination reversal (Figure 2). We therefore use Equation 4 as a sort of "null model." Variations in learning rate, parameter a , therefore indicate deviations that an improved model should account for.

Experienced



Naive

Figure 6. Improvement in discrimination performance across successive discrimination reversals in experimentally experienced (top) and naive (bottom) pigeons (from Davis, 1991, Figure 11). (The rewarded response [S+] is shown by the staggered letters, R [right], L [left], and so forth, at the top; S+ was reinforced with probability $\frac{1}{2}$. Solid line: discrimination performance [$100 * S+ / (S+ + S-)$]; dashed line: a parameter (time constant) computed according to Equation 4 in the text: given $S = [R / (R + L)]$ today and yesterday, and the reinforcement asymptote X ($=0$, for left reinforcement, 1 for right), $a = [S(t+1) - X(t+1)] / [S(t) - X(t+1)]$.)

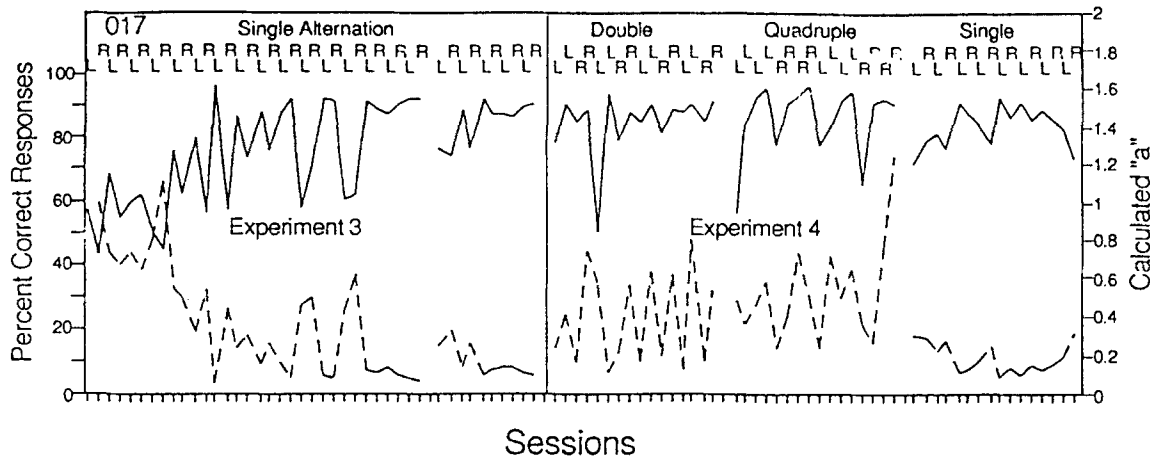


Figure 7. Discrimination-reversal performance of a single pigeon in the two-armed-bandit situation. (Reversal was daily, every 2 days, or every 4 days. S+ was reinforced with probability $1/2$. R = right; L = left. Solid line: discrimination performance each day; dashed line: learning rate each day computed according to Equation 4 [see the legend to Figure 6]. Taken from Davis [1991, Experiments 3 and 4].)

rule and the assumption that the strength of a "silent" response does not change. We therefore preserved both these assumptions in the cumulative-effects (CE) model (Davis, 1991).

Because it is essentially local, model INT-R failed to predict persistent differential effects of previous experience: INT-R always ends up with alternation following R-L-E training even if the R experience (say) is much more extensive. The CE model preserves information from the beginning of the animal's experience in a given situation.

Diminishing returns is a general principle of all learning and both INT-M and INT-R show diminishing returns with respect to behavior. In INT-R, for example, the closer V is to X , the asymptote, the smaller the incremental effect of a reinforcement: $\Delta V = (1 - a)(X - V)$. However, even more important, perhaps, is diminishing returns with respect to reinforcement: The more reinforcements have been obtained (in a given context), the smaller should be the effect of each one. The CE model incorporates diminishing returns with respect to both reinforcement and responses.

These ideas—winner-take-all response rule and constant strength of silent activities, nonlocality, and diminishing returns of response and reinforcement effects—are embodied in the CE model as follows. Each choice is represented by a strength variable, V_i , as before. The model is computed response by response, and the activity with the highest V value is the one to occur, as in INT-R. V values are simply reinforcement probabilities (relative frequencies) in which the reinforcement numerator and response denominator are cumulated from the beginning of the experiment; thus,

$$V_i(t+1) = [R_i(t) + R_i(0)]/[N_i(t) + N_i(0)],$$

$$N_i(0) \geq R_i(0) > 0, \quad (7)$$

where $R_i(t)$ is the total number of reinforcements for response i from the beginning of the experiment through iteration t , and $N_i(t)$ is the total number of times response i has occurred. $R_i(0)$ and $N_i(0)$ are constants representing *initial conditions*—a num-

ber of responses and reinforcements that represent the effect of the animal's prior experience. Because we will always be dealing with symmetrical two-choice situations, we assume that $R_L(0) = R_R(0) = R_0$ and $N_L(0) = N_R(0) = N_0$, so that all our predictions are based on just two free values for initial conditions.

Note that states (values for R and N) in the CE model reflect the model's entire history. It is therefore highly nonlocal in our terms. The predictions of the CE model for the four situations we have already considered—regression, daily reversal, reversal in blocks, and matching—are as follows.

Regression

The type of behavior generated by the CE model during the R-L-E sequence is shown as a response-by-response trajectory in Figure 8. Initial conditions were 200 responses and 200 reinforcements for each of the two choices (note that the origin is at 200,200 and both initial V values are unity). R-only reinforcement lasted for 900 iterations, L-only reinforcement for the next 400 iterations, and extinction thereafter. These three phases are separated by the two vertical lines on the response record.

Because of the winner-take-all response rule and the fact that reinforcement is probabilistic, the two V values (the declining curve) remain close to one another through all three phases of this experiment and cannot be separately identified at the scale of this figure. The reason for the approximate equality of V values (which is useful for analytic purposes, as we show later) is that nonreinforcement reduces the value of the active response until it is less than the value of the silent response, which then occurs, reducing its value, and so on. In the absence of reinforcement, the two responses follow an approximately repeating sequence as the two V values oscillate above and below one another, with the magnitude of the oscillation decreasing as the denominator totals in Equation 7 grow. Each reinforcement provides a transient V -value boost, the magnitude of which de-

creases as the numerator grows. Thus, the average discrepancy between V values decreases with increasing experience.

In Figure 8, the model learns the first discrimination (R) slowly: The response trajectory shows a slowly decreasing slope (i.e., increasing proportion of right responses). It learns the reversal (L) at a similar rate. The model shows the regression effect: The slope of the response trajectory in extinction is clearly less than during the preceding L-only discrimination. Moreover, with this particular history, the slope of the trajectory in extinction is approximately 0.7, with more responses going to the right, which is the longer-trained alternative. Thus, the CE model is not constrained to predict indifference in extinction and is sensitive in a plausible way to the amount of experience with each of the two choices. The effect is a regression, in the sense that in extinction after a reversal the model will usually revert to a less extreme preference. The CE model also predicts that under most conditions, reconditioning after extinction will be faster than original conditioning, a finding that poses difficulties for all performance models.

There is a straightforward geometrical interpretation of this regression property of the CE model, which is illustrated schematically in Figure 9 for the simple case where initial conditions are negligible.⁶ The figure shows the cumulated total of reinforcements for left and right responses plotted against the cumulated total number of responses. The two filled circles show the values of R_L versus N_L (lower circle) and R_R versus N_R (upper circle) at the end of the L-only reinforcement phase in Figure 8. The R-only phase was longer than the L-only phase (so that $R_R > R_L$ and $N_R > N_L$), but, because of the winner-take-all rule, the overall reinforcement probability, R/N , must always be approximately equal for both responses (see the two declin-

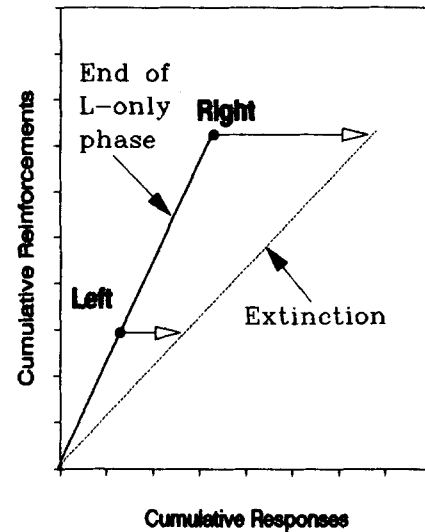


Figure 9. Schematic illustration of the cumulative-effects model predictions in extinction following right-left discrimination reversal. (L = left. See the text for details.)

ing V -value curves in Figure 8). Thus, the two filled circles representing total R versus total N for left and right at the end of the L-only phase (filled circles) must lie on the same straight line through the origin, whose slope (if initial conditions are negligible) is approximately equal to R/N , the heavy line in Figure 9. In extinction, responses (arrows), but not reinforcements, are added to both these two points, which are therefore displaced along the x -axis. Moreover, because of the winner-take-all response rule, responses occur so as to keep the two V values (overall reinforcement probabilities) equal. Thus, the two points representing right and left choices must always remain on the same line through the origin, although the slope of this line will continuously decrease as more responses occur. The dashed line in the figure shows the state of the model after some time in extinction. Because of the geometry of the situation, more responses must be made to the right choice, and the ratio of right to left responses at any point during extinction will be approximately equal to the ratio of right to left reinforcements during training (including initial conditions).

There are three things to notice about the CE extinction prediction. First, it is consistent with the experimental results of Myerson and Hale (1988), which were described earlier. They found that the ratio of left to right responses in extinction continued to match the ratio of reinforcement rates after a history in which the proportion of reinforcements received for left and right was maintained constant in each experimental condition. Second, the predictions are also in general agreement with many experimental results showing that amount of training has an effect on preference in extinction. Third, the CE model predictions suggest an explanation for the sometimes extreme indi-

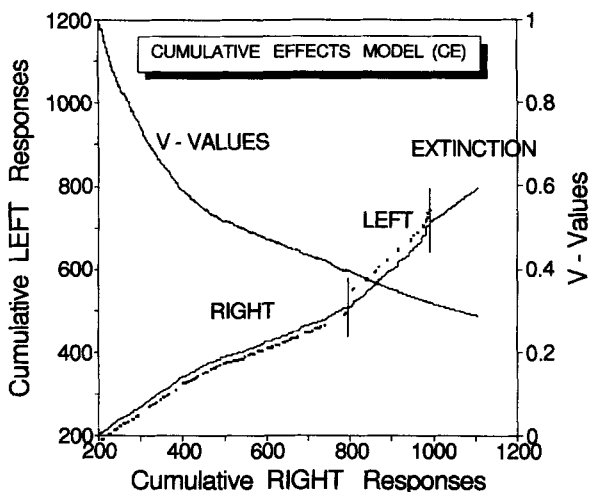


Figure 8. A prediction of the cumulative-effects model for discrimination learning (right-only reinforced) and reversal (left-only reinforced) followed by extinction in the two-armed-bandit situation. (Reinforcement probability was $1/2$ and initial conditions were 200,200 [reinforcers and responses for each choice]. Declining curves: V values for each choice, computed according to Equation 7 in the text; rising curve: cumulative right responses plotted against cumulative left responses. Dots below [right] or above [left] the line indicate reinforcements.)

⁶ Geometric (or algebraic) analysis is possible only when V values are approximately equal, which is true under many, but not all, conditions. We restrict formal analyses to cases where this condition is satisfied.

vidual differences in preference in extinction that are shown by animals with different long-term reinforcement histories. Finally, the CE model allows for a dissociation between preference, which is determined solely by V value, and persistence in extinction, which is determined by the number of reinforcements that have been received for a given choice. Because short-term preference and persistence do not always covary empirically, this dissociation is a useful property.

Successive Daily Reversal (SDR)

The CE model permits improvement in discrimination performance across successive reversals. A typical choice-trajectory simulation, with the same initial conditions and payoff probabilities as in Figure 8, is shown in Figure 10. It is easy to see that the model tends to switch preference more rapidly in later reversals: The proportion of correct responses at the end of each discrimination is shown by the filled squares (right-hand y -axis). The improvement across reversals is smoother if reversals occur after a fixed number of reinforcements rather than after a fixed number of iterations (see Figures 11 and 12).

The effect of initial conditions on the pattern shown in Figure 10 is rather complex. With less initial experience (10 responses and reinforcers for both choices, 10,10, instead of 200,200) the excursions in performance from one reversal to the next are more extreme and the pattern across reversals is less clear. With more initial experience (1000,1000) excursions are reduced, and there is still improvement, but it takes more reversals to become apparent. The CE model is not able to duplicate the much impaired performance on the first few reversals shown by the naive animals in Figure 6.

A geometric interpretation of the SDR improvement shown

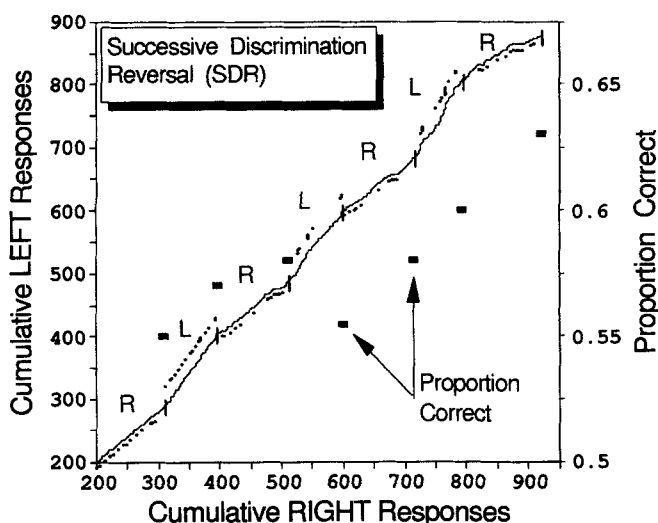


Figure 10. Successive discrimination reversal predictions of the cumulative-effects model. (Initial conditions: 200,200 and reinforcement probability was $1/8$. Each discrimination phase lasts for 200 responses. Filled squares are the proportion of correct responses at the end of each reversal [right-hand y -axis]. Dots below [right] or above [left] the line indicate reinforcements. R = right; L = left.)

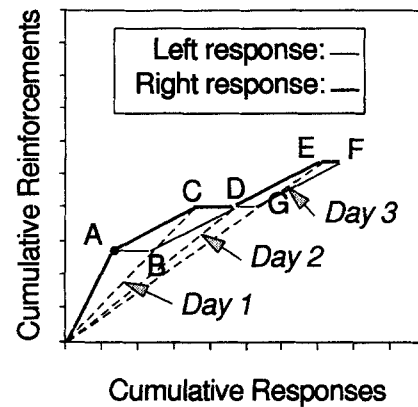


Figure 11. Schematic depiction of the cumulative-effects model predictions for successive discrimination reversal; the initial 4 days—right, left, right, left—following continuous reinforcement training are illustrated. (See the text for details.)

by the CE model is shown in Figure 11. The figure shows 4 days of an R-L-R-L reversal sequence following training with equal numbers of responses and reinforcements for each of the two choices (CRF). As in Figure 9, the axes are cumulative responses and cumulative reinforcements. At the end of CRF training, cumulative responses and reinforcements are the same for both choices: point A in the figure. During the first discrimination session, only right responses are reinforced. By the end of that session, right responses have followed a trajectory of constant slope (because reinforcement probability, R/N , set by the schedule is constant) for a constant number of reinforcements (vertical displacement) arriving at point C. Over the same time, a number of unreinforced left responses have been made, such that at the end of the session the V values (total reinforcement probability) for left and right are the same. Thus, the left responses follow a horizontal trajectory terminating at point B such that at the end of this first session the points representing total reinforcements plotted against total responses lie on the same line through the origin: the dashed line labeled *Day 1*.

Because the reinforced response is paid off with constant probability on these two-armed-bandit schedules and sessions are a fixed number of reinforcers long, the number of correct ($S+$) responses is the same each day. Thus, the level of daily discrimination performance is determined entirely by the number of erroneous ($S-$) responses. The left-response errors the first day correspond to light-line segment AB. Left responses are reinforced on the second day (line segment BD). The next opportunity for left errors is Day 3, represented by line segment DG, which is shorter, indicating fewer errors, than segment AB on Day 1. A similar process operates for right responses (heavy-line segments): There are more right errors on Day 2 (line segment CD) than on Day 4 (line segment EF). Thus, the CE model usually implies that discrimination performance will improve across successive discrimination reversals.

There are a number of alternative accounts for improvement across successive reversals. Some of these, such as reversal learning "set," are not well defined. Others, such as the idea that the subject uses the outcome of the first response each day as a cue

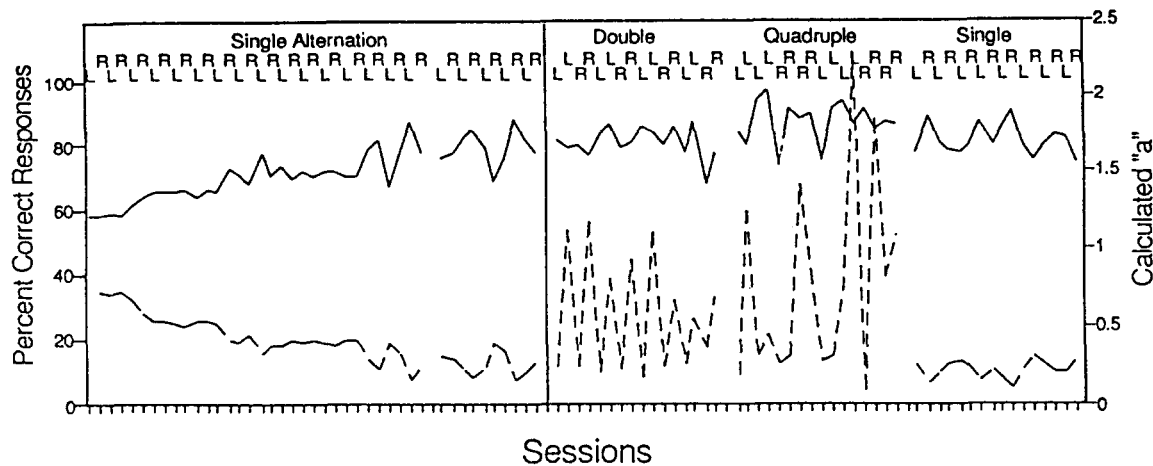


Figure 12. The cumulative-effects model simulation of the data shown in Figure 7. Initial conditions: 1000,2000. (R = right; L = left. See the text for details.)

to the identity of S+ on that day (the "win-stay, lose-shift" hypothesis, WSLS), are more complex than the CE model and no more general. The WSLS hypothesis is more complex because it implies some kind of associative process that identifies predictive stimuli, plus some other process that switches preference on the basis of that identification. Without some quantitative form, this hypothesis cannot deal with the different effects of daily reversal versus reversal less frequently, nor can it deal easily with reversal under intermittent reinforcement, as in the Davis data. However, the main argument for the CE model is parsimony: It accounts for many of the features of discrimination reversal by means of the same, simple process that works for the initial learning.

Reversal in Blocks

Figure 12 shows an extensive CE-model simulation of the block-reversal data in Figure 7. The conditions for the simulation were those used by Davis (1991): A daily session was defined as having 48 reinforcers, and the S+ reinforcement probability was $\frac{1}{8}$. The initial conditions for the simulation were 1,000 reinforcers and 2,000 responses.

The simulation duplicates the three main features of the data in Figure 7:

1. Improvement in reversal performance across SDRs. This is shown both by daily improvement in percentage correct responses (solid line) and by the increased learning rate (reduced value of a , computed from Equation 4, dashed line) across the series of daily reversals (Single Alternation) in the left-hand panel and on the far right of the right-hand panel.

2. Change in learning rate (parameter a in Equation 4) as a function of frequency of reversal: a is lowest in daily reversal and highest when reversal is every 4 days.

3. Change in learning rate within and between blocks in the 2- and 4-day-reversal conditions: the computed a value increases within each block and then decreases between blocks.

These features depend to some extent on initial conditions. Improvement across successive discrimination reversals seems

to require initial conditions with large absolute values for R and N and requires $V(0)$ to be in the vicinity of 0.5 (e.g., 1000,2000, as in Figure 12). When the absolute values are even larger (e.g., 5000,10000), improvement is slower and may not reach as high an asymptote. When the initial conditions are small (e.g., 1,2), discrimination learning is rapid on Day 1 but much poorer the next day (first reversal), just as with our naive pigeons. Unlike the data, however, performance does not improve much on subsequent reversals, because performance gets "out of phase" with the schedule (on Day $N + 1$ the model perseverates in behavior appropriate to Day N). It is possible to estimate initial conditions from the data, but the estimate is not very sensitive. Two methods are discussed in Appendix C.

Matching

As we saw earlier, matching entails equal reinforcement probabilities for the two alternatives: $R_i/N_i = R_j/N_j$ for a given pair, where these quantities are computed over a limited period; occasionally these quantities are computed over the whole time a given pair of VI schedules is in effect, but usually they are computed over the latter part of this time, when preference has stabilized. The CE model also implies that reinforcement probabilities will be equated (Equation 7). The difference is the period over which the quantities are computed, which, for the CE model, is the entire experiment, including initial conditions. This difference makes less of a difference than might be expected. For example, in one 1,000-iteration simulation with VI set-up probabilities of 0.05 and 0.1 during each iteration and initial conditions 100,100, the ratios were $R_R/R_L = 2.5$ and $N_R/N_L = 1.56$ for the whole 1,000 iterations and $R_R/R_L = 2.4$ and $N_R/N_L = 1.73$ for the last 500 iterations. Because the response ratio is less extreme (closer to unity) than the reinforcement ratio, both cases are examples of undermatching, which is the usual systematic deviation from matching (Baum, 1979; Davison & McCarthy, 1988). However, the difference between the two averaging periods is relatively small. As additional conditions are added (so that R and N both increase for both

choices), with the usual care taken that both choices are equally reinforced overall, the degree of undermatching decreases further.

The typical finding that undermatching decreases with increasing exposure to a given pair of schedules follows at once from Equation 7: $V_i(t+1) = [R_i(t) + R_i(0)]/[N_i(t) + N_i(0)]$, because the contributions of the constant terms, $R_i(0)$ and $N_i(0)$, are reduced as $R_i(t)$ and $N_i(t)$ increase and the value of each V_i converges on R_i/N_i . This prediction is consistent with many published (e.g., Todorov, Castro, Hanna, de Sa, & Barreto, 1983) and unpublished results. However, the prediction that undermatching decreases with increasing R and N totals is not consistent with another finding of Todorov et al. They evaluated undermatching by means of the standard power-law equation: $x/y = k[R(x)/R(y)]^w$, where x and y are response rates and $R(x)$ and $R(y)$ are obtained reinforcement rates, averaged across the last five sessions of each condition. *Undermatching* is defined as a value for exponent w less than unity. Todorov et al. reported that the value of w computed from conditions 1–5, 1–6, 1–7, and so forth, decreased; that is, undermatching increased through the experiment, even though undermatching decreased (i.e., matching was better approximated) during later sessions of a given condition.

We have applied the CE model in detail to Herrnstein's (1961) original concurrent VI-VI experiment. The result is close-to-perfect matching of response ratios to reinforcement ratios.⁷ The results of this simulation are compared with Herrnstein's data in Figure 13, and they match the original data almost point for point. This outcome is almost independent of initial conditions as long as they are not extremely large; if they are smaller than 1000,2000, the points lie closer to the diagonal.

The fact that matching is readily predicted by the CE model is interesting because it shows matching may be produced by a nonlocal model. Melioration (Herrnstein & Vaughan, 1980) as well as other, less popular, models for matching, such as ratio invariance (Horner & Staddon, 1987) and the kinetic model (Myerson & Miezins, 1980), all assume that matching must be a local process.

IV. Discussion

We have discussed two types of choice model, state and performance models, and a model property, locality. State and performance models differ in the relation between the state of the model and the measurable behavioral properties. The state of performance models is isomorphic with easily measured behavioral properties such as rates or probabilities. Identifying the state of a state model requires either more extensive knowledge of the organism's past history or a series of preliminary "setting" operations. Local models do not indefinitely preserve information about remote past history.

We have shown that the molar integrator, INT-M, a local, performance model that provides a pretty good description of chronic individual-animal data from choice discrimination-reversal experiments, nevertheless fails to capture relatively simple historical effects such as regression in extinction following a single reversal. It also fails to capture more complex effects such as improvement in speed of learning across reversals and a de-

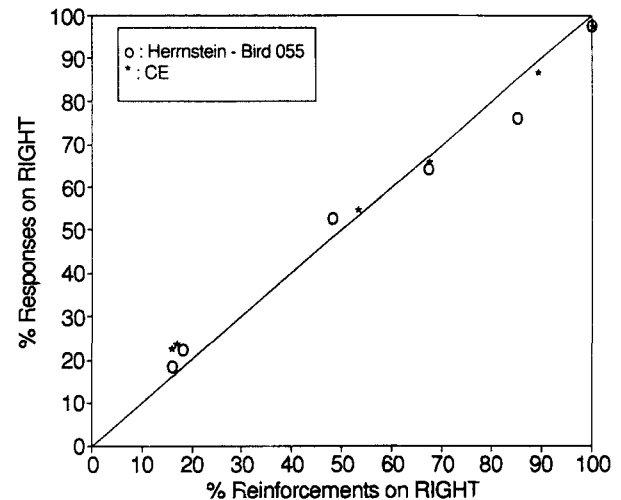


Figure 13. Matching on concurrent variable-interval (VI) schedules. (Points show the average proportion of right responses plotted against the proportion of reinforcers obtained for right responses for a range of VI-VI pairs, taken from Herrnstein (1961). Open circles: Herrnstein data; asterisks: cumulative-effects[CE] model predictions. The CE simulation assumed one response per second and initial conditions 1000,2000. Points are averages from the last five sessions under each condition in both cases.)

pendence of learning rate on the frequency of reversals. INT-M requires advance specification of the asymptote for learning each day, so it cannot predict the ubiquitous matching relation. A response-by-response, local state model that is based on the same process, INT-R, does better in predicting the regression effect, predicts undermatching, but also fails to generate performance improvement across reversals or dependence on reversal frequency.

These failures were an indication to us that an adequate model for recurrent choice is likely to be nonlocal. The assumption that response tendencies compete nonlinearly, according to something like the WTA rule, seems to be essential to the prediction of regression effects following discrimination reversal and faster reconditioning after extinction. We therefore devised a nonlocal, state model, the cumulative effects (CE) model, which retains the winner-take-all response rule, as a simple way to test these conjectures. We have shown that the CE model provides a qualitative explanation for regression effects, faster reconditioning after extinction, matching, improvement across reversals, and dependence of learning rate on reversal frequency.

The CE model can also shed light on other effects in the literature. Williams and Royalty (1989) carried out a complex experiment to test an implication of one version of Herrnstein and Vaughan's (1980) melioration hypothesis. Their procedure

⁷ The CE model cannot deal with the *changeover delay* (COD) procedure used to prevent frequent switching between keys in many choice experiments. This is partly a reflection of the absence of time as a variable in the model. To simulate Herrnstein (1961), we therefore had to set the COD to zero.

allowed pigeons to be trained simultaneously on two concurrent VI-VI choice schedules. For example, under one set of stimulus conditions, the animals had to choose between a VI 20-s and a VI 120-s schedule. Under an alternating set of stimulus conditions, they had to choose between a VI 60-s and a VI 80-s schedule. We label these four schedules A, B, C, and D, in order of reinforcement frequency, A = VI 20, and so forth.

Animals match reinforcement probabilities under both these training conditions, A versus D and B versus C. However, the absolute reinforcement probability at which a match is struck will obviously be higher under the VI 20-VI 120 condition (A vs. D) than under the VI 60-VI 80 (B vs. C) condition; the equilibrium matching probability is limited by the richer of the two VI schedules. The critical test for melioration in this experiment was to pit the stimulus for VI 60 s (B) against the stimulus for VI 120 s (D) in extinction. One view of melioration predicts that because the reinforcement probability associated with stimulus D (paired with A in training) is higher than that associated with stimulus B, the pigeons should prefer D in a test, despite its lower real reinforcement rate. This was the interpretation tested by Williams and Ryalty (1989).

The CE-model prediction is illustrated schematically in Figure 14. The points representing the numbers of responses and reinforcers on the concurrent VI 20-VI 120 schedule (A vs. D) lie on a steep line through the origin. They lie on the same line because of the matching of reinforcement probabilities implied by the winner-take-all response rule. Point A is above point D because the absolute number of reinforcements obtained by responding to A is higher than the number obtained by responding to D, because the animal spends more time responding on VI 20 than on VI 120. The points representing the numbers of responses and reinforcers on the concurrent VI 60-VI 80 schedule (B vs. C) lie on a shallower line; point B is above point C for the same reason that point A is above point D on the steeper line: In both cases, the higher point on the line corresponds to the "richer" VI schedule.

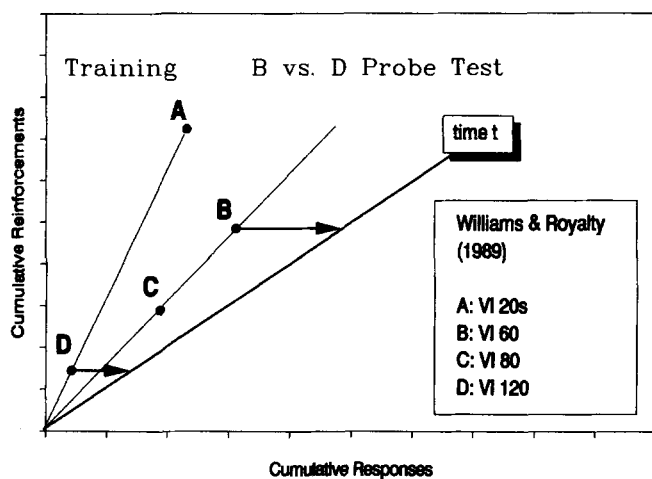


Figure 14. Schematic depiction of the predictions of the cumulative-effects model of an experiment by Williams and Ryalty (1989). (Initial conditions are assumed to be negligible [$\approx 0,0$]. VI = variable interval. See the text for details.)

In extinction, stimuli B and D are pitted against one another. As extinction responses occur, each point is displaced along the x-axis (arrows), with both points remaining on a line through the origin because of the winner-take-all response rule. The situation after time t in extinction is shown by the heavy line. It is clear that the prediction of the CE model depends on the period over which data are averaged in extinction. Initially, the model "prefers" alternative D, because its V value (reinforcement probability) is higher than the V value for alternative B. However, as D extinction responses accumulate, the two V values soon become equal and both B and D responses occur thereafter so as to keep them equal (i.e., on the same line through the origin). As the figure shows, at time t and thereafter, more responses will be made to alternative B than to alternative D. These predictions are changed quantitatively, but not qualitatively, if the values for initial conditions are significant.

Williams and Ryalty (1989) presented average data from an entire extinction session. If this time is large in our terms, then the CE model implies a preference for alternative B over D, which is what they found. Our analysis suggests that the time over which the B versus D preference is measured will be critical to the direction of preference; the best form of data presentation from our point of view would obviously be the response trajectory in extinction, because the clearest prediction from the CE model is a shift in preference as more extinction responses accumulate.

A very similar analysis allows the CE model to account for recent data on "transitive inference" (TI) in pigeons (Fersen, Wynne, Delius, & Staddon, 1991; see also Couvillon & Bitterman, 1992; Wynne, Fersen, & Staddon, 1992). The TI effect is based on tests after training with a series of four successively presented, overlapping, simultaneous discriminations: A+B-, B+C-, C+D-, and D+E-. The TI effect probably depends on the order in which these four discriminations are trained and the amount and type of training on each, but the full picture is not yet known. In the Fersen et al. experiment, the order of exposure was random and all pairs got equal numbers of reinforcements. There are three critical findings from this experiment and preceding TI studies with humans and other species (see Fersen et al. for references):

1. At the end of training, performance on A+B- and D+E- discriminations is much better than on the ambiguous pairs, B+C- and C+D- (this is known as the *end-anchor* effect).
2. However, performance on D+E- is almost invariably better than performance on A+B-, even though D is unrewarded when paired with C and A is always rewarded.
3. In transfer tests in extinction, B is usually preferred to D, even though B and D are never paired in training, and B is often more preferred to D than to C, with which it is paired in training. This B > D transfer is the TI effect.

A version of the CE model using a matching response rule was tested by Wynne et al. (1992) and predicted correctly the rank order of discrimination performance on the four training effects; this version did not predict the superiority of B versus D to B versus C that Fersen et al. (1991) found, however. The way that the present version, with a winner-take-all response rule, accounts for TI in this situation can be represented geometrically for the case where the four discriminations are trained in blocks: first A+B-, then B+C-, C+D-, and D+E- (an experi-

ment with this design, which found a strong TI effect in pigeons, was recently reported by Steirn & Zentall, 1991). Discrimination phases 1, 2, and 3 are depicted at the top of Figure 15. Phases 3 and 4 in the bottom panel. We assume that the model has equal experience with the five stimuli at the beginning of discrimination training so that initial conditions are the same. In this figure, and those that follow, we assume a high value for the initial V s (i.e., a high R/N ratio), which is consistent with the usual practice of training animals on continuous reinforcement before shifting to an intermittent schedule. Points in the figure labeled with lowercase letters indicate the values of R (cumulated reinforcement) and N (cumulated responses) at the beginning of training on a given discrimination; capital letters indicate these values at the end of a discrimination. Thus, the points where a , b , c , d , and e first appear correspond to the initial conditions and have the same coordinates.

Look first at the diagram labeled $A+B-$. Stimuli A and B begin with the same strengths (initial conditions: point labeled a, b). We assume that responses to A are reinforced according to a probabilistic schedule. Thus, at the end of a fixed number of reinforcements, the point representing A will have traveled fixed vertical and horizontal distances, represented by line segment aA . The number of unreinforced responses to $B-$ is represented by horizontal line segment bB , whose length is such that at the end of the $A+B-$ discrimination the two V values, for A

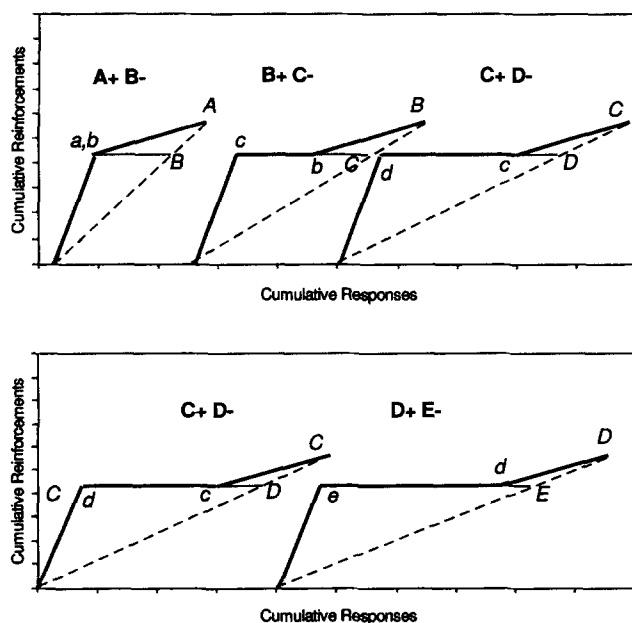


Figure 15. Schematic depiction of the cumulative-effects-model predictions for the transitive-inference effect in the same coordinates as Figure 14. (The model is trained first with stimulus A associated with probabilistic reinforcement and stimulus B with extinction [$A+B-$], then $B+C-$ followed by $C+D-$ and $D+E-$. $A+B-$, $B+C-$, and $C+D-$ are shown in the top panel, and $C+D-$ [repeated] and $D+E-$ are shown in the bottom panel. Lowercase letters indicate the R [reinforcements] versus N [responses] coordinates for a given stimulus at the beginning of the discrimination, whereas capital letters indicate the R versus N coordinates at the end. Note that each diagram begins at 0 on the x -axis. See the text for details.)

and B , will be the same; that is, A and B will lie on the same line through the origin.

At the beginning of the second discrimination, $B+C-$, C begins at the initial-conditions point, labeled c , as before. However, B is at the same point it was at the end of the $A+B-$ discrimination, which is labeled b at the beginning of the $B+C-$ discrimination. As the reinforced alternative, responses to B again follow the fixed-slope trajectory labeled bB (the same as aA in the $A+B-$ discrimination). Error responses to C are represented by the horizontal line segment cC . This same pattern is repeated in the other two discriminations, $C+D-$ and $D+E-$: The previously negative stimulus (e.g., C) begins each time at a larger N coordinate; the new stimulus begins at the initial conditions. Thus, each new discrimination is learned more slowly ($bB < cC < dD < eE$), which has been reported in at least one of these experiments.

The TI effect emerges at once from these diagrams. Look first at the capital letters showing the states of the four reinforced stimuli at the end of discrimination training. All four reinforced stimuli have the same R coordinate at the end of training (because all are reinforced equally often); the relative values of the stimuli depend only on the differences in N coordinates, which correspond to slope differences. Thus, the stimulus values are linearly ranked, A has the steepest slope, B the next steepest slope, and so on: $A > B > C > D > E$. Hence, in any novel comparison, such as B versus D , the leftmost stimulus will be preferred, and the preference for B over D will be greater than the preference for B over C , which are the two TI effects.

These diagrams show the $A+B-$ end-anchor effect (fewest errors on the $A+B-$ discrimination) but not the $D+E-$ effect. This effect and all the others are likely to depend not only on initial conditions but also on the way that discriminations are intermixed during training. Exploring these possibilities requires more comprehensive simulations, which we expect to report on elsewhere. Here we simply show that the CE model is compatible with the TI effect.

Limitations

The state variables of the CE model are simply totals and contain no explicit representation of the order of events. It might appear, therefore, that the model is insensitive to the order of experimental conditions, which is inconsistent with several well-known results. However, even though order is not represented explicitly in the state variables of the model, the behavior of the model is in fact highly sensitive to the order of experimental conditions. For example, in one set of simulations we looked at the state of the model, with initial conditions 1, 2, after 20 sessions of L-only reinforcement followed by 5 sessions of R-only (20L, 5R), or the reverse (5R, 20L). The reinforcement totals were the same in both cases, but the response totals were quite different, on the order of 1,800 (N_R) and 15,000 (N_L) in the first case and 6,000 (N_R) and 7,000 (N_L) in the second. Thus, the state of the model, hence its future behavior, is quite different after L,R training versus R,L training.

Formal analysis shows that this order sensitivity is because the assumption that V values are approximately equal is violated. With this training history, and many others, V values can

diverge substantially. When V values are not equal, simulation is the only way to derive predictions from the model.

A specific failure of the CE model is its inability to account for the difference in the way that naive and experienced animals improve across successive discrimination reversals, illustrated in Figure 6. The obvious possibility is that this difference can be duplicated by adjusting initial conditions, but we have not been able to duplicate this effect except by giving explicit extra weight to R and N increments on the very first discrimination session. Evidently, the very first exposure to a discrimination is especially important in ways that are not captured by any of the models considered here.

The CE model is deterministic, whereas the theoretical fashion in learning has always been for stochastic models. Because it is deterministic, it makes some wrong predictions at the molecular level. For example, it predicts that switches from one choice alternative to another only occur after one or more unreinforced responses and never after reinforcement. This is contrary to fact: In unpublished experiments both Machado and Horner in this laboratory have found idiosyncratic, but reliable, postreinforcement switching on probabilistic choice schedules. The failure to accommodate postreinforcement switching could be partially remedied by adding a stochastic element. However, because such an addition would be entirely ad hoc, would complicate analysis, and would add nothing to our understanding of basic learning processes, it seems better to live with this known limitation until we understand the deterministic model better.

The lack of any kind of discrimination threshold might also be thought to be a problem. In its simplest form, the CE model assumes that the highest V value wins, irrespective of the magnitude of the difference between the highest and next-highest V value. This seems improbable, as does the assumption that the values of R and N can rise without limit. These are reasonable objections, which could be answered by rescaling R and N , changing the V computation from division to subtraction, and adding a stochastic assumption, but seemed to us secondary to the main issue, which is the empirical reach of the model, where it has more serious limitations that need to be resolved before secondary questions need be addressed.

The major general limitation of all these models is their failure to incorporate real time: The models are dynamic in the sense that they deal with changes in behavior, but they are not real-time models. They cannot account, therefore, for inter-trial-interval SDR effects (Williams, 1976), effects on absolute response rate, or effects that depend on temporal discrimination.

One might also suspect that the CE model will have difficulty with other effects that seem to be dependent on rates of occurrence, such as the partial reinforcement extinction effect (PREE), the successive negative contrast effect (SNCE), and the overtraining reversal effect (ORE). These effects were more studied in the heyday of Hull-Spence learning theory (see the review in Mackintosh, 1974), and most experiments used mazes and runways and trial-by-trial procedures rather than the free-operant procedures we deal with here. Moreover, the evidence for some of these effects in free-operant experiments is often equivocal. The CE model can deal with some aspects of these effects but has difficulty with others.

The PREE is the greater persistence of responding following intermittent reinforcement. It is easy to see from a geometric diagram like those in Figures 9, 11, 15, and 16 that whether we equate the number of preextinction reinforcements, iterations, or responses, extinction will be more rapid (there will be fewer responses at any point in extinction) following intermittent than following continuous reinforcement when the two are compared in a choice context. Many PREE experiments fail to equate numbers of reinforcements, however, but rather train animals in the continuous and partial groups to a performance criterion. It is easy to show that the partial animals will take more reinforcements than the continuous animals to achieve the same performance level, in which case the CE model will often predict more extinction responses following partial training.

In fact, the PREE seems to occur rather rarely in free-operant experiments. We have not been able to find a choice experiment of the appropriate sort, but Nevin (1988) has reviewed a number of similar experiments with multiple schedules (i.e., successive rather than simultaneous discriminations). For example, in one experiment (Nevin, 1979, Fig. 4.1), pigeons were trained on multiple VI 28-s-VI 86-s, VI 86-s-VI 360-s, or VI 28-s-VI 360-s schedules. In extinction, responding in the higher reinforcement-rate component always extinguished more slowly than responding in the lower reinforcement-rate component: the opposite of the standard PREE but consistent with CE-model predictions.

The CE model implies that the most important variable determining resistance to extinction in a choice situation is the number of reinforcers received for each alternative. Nevin (1988) noted that when it is possible to compare the effects of number of reinforcers, resistance to extinction increases with the number of reinforcers received in training.

The SNCE is the name given to the crossover in responding that occurs when reinforcement conditions are reduced either to zero or to a low value, following either a relatively rich or a relatively lean schedule: Responding declines to zero faster

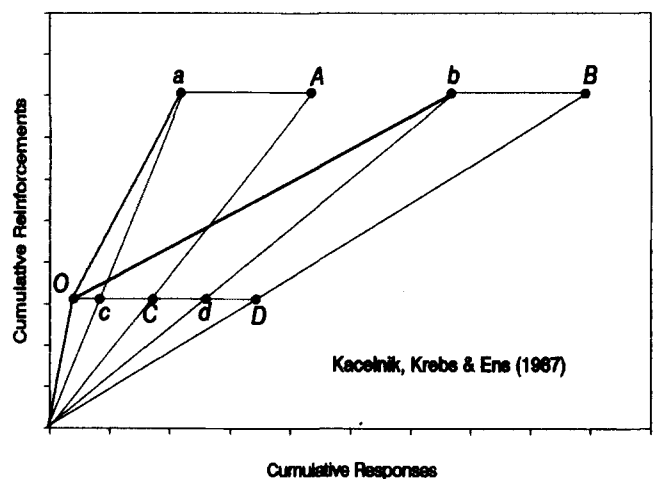


Figure 16. Schematic illustration of a cumulative-effects analysis of the successive negative contrast effect in an experiment by Kacelnik, Krebs, & Ens (1987). (See the text for details.)

after the rich schedule than after the lean schedule. The CE model may be able to account for some instances of the SNCE, particularly those in which more reinforcements overall are received for the leaner schedule, but it cannot account for cases where the number of reinforcements is equated. The way the argument works can be illustrated by a recent experiment by Kacelnik, Krebs, and Ens (1987). Starlings chose between two "foraging patches" in which food was dispensed according to either a rich or a lean probabilistic schedule. There were two comparison conditions; in both, the lean schedule was $p(F) = 0.08$. In one comparison, rich was $p(F) = 0.25$; in the other, rich was $p(F) = 0.75$. After at least 100 reinforcers had been obtained on both rich schedules, reinforcement on the rich patch was withdrawn. The question was, would responding on the rich side by the animals trained with the 0.25, 0.08 pair decline more slowly in extinction than responding on the rich side by the animals trained on the 0.75, 0.08 pair? Kacelnik et al. found an SNCE-like effect: "the number of responses after the drop to zero before the first switch to the stable side was larger than in the 0.25–0.00 treatment than in the 0.75–0.00 treatment for all three birds" (p. 71).

A schematic CE analysis of this situation is shown in Figure 16, with the lean schedule set to $p(F) = 0$ (rather than to 0.08) for simplicity. Point O is the initial conditions for both comparisons: 0 versus 0.75 (represented by polygon $OaAC$) and 0 versus 0.25 (represented by polygon $ObBD$). Point a is the number of responses and reinforcements for the 0.75 schedule after a fixed number of reinforcements in training; point c is the 0 schedule at the end of training. The comparable two points for the 0 versus 0.25 schedule are b and d . In extinction, after a fixed number of responses to each of the rich schedules, the 0.75 schedule has moved from a to A and the 0.25 schedule the same distance, from b to B . In the 0 versus 0.75 comparison, the number of responses to the 0 choice is line segment cC ; in the 0 versus 0.25 comparison, the number of responses to the 0 choice is dD . With the initial conditions shown, cC is the same length as dD , independent of the absolute $p(F)$ values; that is, the model makes as many responses to the lean schedule in the 0.25 versus 0 comparison as in the 0.75 versus 0 comparison. This is half way to the classical SNCE effect: At least there is not more responding to the rich schedule than to the lean schedule, but there is no actual crossover. The CE model is only partly compatible with the SNCE under these conditions.

The ORE is the fact that animals trained to go left (say) will learn the opposite discrimination (go right) more rapidly if the left training is more extensive, which is a paradoxical result from the point of view of a simple strength theory of instrumental conditioning: More training should mean more strength for the left response, which should mean slower learning of the opposite, right response. Not all who have looked for the ORE have found it, however, and the effect sometimes seems to be biphasic: Over a moderate range, more initial training means slower reversal; but after very much initial training, reversal performance improves. Mackintosh (1974) writes: "All reviewers have been forcibly impressed by the inconsistency of the effect and its reluctance to submit to any simple analysis" (p. 603). Comparing CE-model predictions with published data is obviously difficult here because of procedural differences, the absence of detailed performance information (other than

"trials to criterion," for example), and the inconsistency of the effect. Nevertheless, it is of some interest that under many conditions the CE model predicts improved reversal performance after more extensive initial training.

The prediction can be represented schematically. Figure 17 shows the formation of a discrimination followed by a reversal. The model begins with identical initial conditions for each choice at point A . Then right (say) is reinforced according to a probabilistic schedule; left gets no reinforcements. After a given training period, the state of right is represented by point B , which represents the addition of responses and reinforcers in a fixed proportion represented by line segment AB . Over the same training period, the model makes the number of left error responses represented by line segment AD , which lies on the dashed matching line OB . The discrimination is then reversed for the same number of reinforcers, so the state of the left response is represented by point C , such that line segment DC has the same length and slope as line segment AB . Over the same period, BC represents the right error responses. The thing to attend to is the length of BC as segment AB increases with training. Although it is difficult to show in a diagram because of the magnitudes involved, it should be obvious that as AB increases with training in the direction of the upper arrow, the matching line, OB , rotates in the direction of the lower arrow so that its slope approaches closer and closer to the slope of the schedule line AB . In other words, after sufficiently protracted training, discrimination reversal will entail negligible errors.

V. Conclusion

Two traditional, local learning models fail to capture basic phenomena of serial reversal learning. A simple, nonlocal model, the cumulative-effects (CE) model, is much more successful. The CE model is also consistent with several experiments on extinction after concurrent discrimination. The model shows remarkably complex behavior that resembles the general features of the behavior of pigeons and rats under several different choice-procedure histories: a range of situations

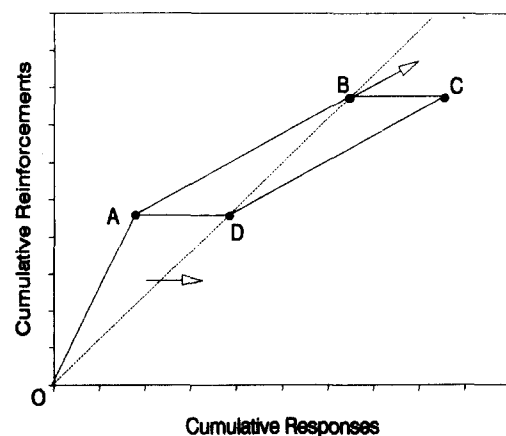


Figure 17. Schematic illustration of a cumulative-effects analysis of the overtraining reversal effect.

much wider than that encompassed by any other operant-learning model of which we are aware. We conclude that the process that drives behavior in free-operant choice experiments with animals is almost certainly nonlocal: Remote events, preceding the current conditions of reinforcement, continue to have an effect on behavior. The pigeon, like the elephant, does not forget.

Three features of the CE model seem to be critical to these predictions: winner-take-all competition among parallel, independent processes; unchanging strength of unexpressed (silent) activities; and state variables that retain information from the beginning of the experiment. The fact that the CE model represents the state of a given response by two variables, R and N, rather than by just a single strength variable is also essential.

The CE model is response level and therefore cannot predict properties of choice related to time; it may not be able to handle some effects of intermittent reinforcement. However, these failures are perhaps better regarded as guides to future theoretical development than fatal flaws in the approach.

The traditional theoretical approach in many areas of psychology emphasizes quantitatively precise tests of analytically tractable models in limited, fixed experimental situations, usually with groups of subjects. This approach has carried over to studies of learning, even though the essence of learning is individual differences and adaptation to variable conditions. Instead of seeking quantitative precision in a limited context, we therefore argue in this article for an alternative approach: qualitative tests of conceptually simple models (which may be relatively tractable analytically, like the CE model, but need not be) of the behavior of individual animals subjected to a relatively wide range of experimental situations. By comparing the strengths and weaknesses of a variety of models under a variety of conditions, we hope to assemble more and more comprehensive models that may eventually approximate the behavior of our subjects under the widest possible range of conditions.

The frequent assumption that the study of learning requires between-group experiments is not correct. The learning of individuals may be understood by exploring simple models for their behavior in a series of different learning tasks.

It may fairly be said of reinforcement history as it is about the weather that everyone talks about it, but no one does anything about it. B. F. Skinner argued for reinforcement history as the ultimate cause of everything (see, for example, Rachlin, 1991), but in practice most workers in this field (ourselves included) have focused on the current conditions of reinforcement. The emphasis in recent years has been almost exclusively on reversible phenomena, that is, on steady-state analysis of performances apparently determined entirely by current conditions (cf. comments in the review by Dow & Lea, 1987). Unfortunately, a reversible *phenomenon* may or may not correspond to a reversible *state* of the organism: The "same" behavior does not imply the same state. Recall that the CE model predicted matching both early and late in Herrnstein's (1961) experiment, but the state of the model early and late in training, and therefore its response to new conditions (like extinction, for example), is very different. Who can doubt that the same is true of the pigeon?

In this article, we have tried to show that theoretical models

provide a way to get at remote historical effects: at persistent effects of experimental conditions preceding the current condition—at the state that underlies the behavior of the individual organism. This attempt is only a beginning. Nevertheless, given the relative ease with which predictions may now be made from simple, nonlinear models it is not difficult to see how this approach may be extended to a much wider range of models and experimental situations. We hope that the work we have described will be only the first of many attempts to explore these new directions for learning theory.

References

- Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, 32, 269–281.
- Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.
- Couvillon, P. A., & Bitterman, M. E. (1992). A conventional conditioning analysis of "transitive inference" in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 18, 308–310.
- Davis, D. G. S. (1991). *Probabilistic choice: Empirical studies and mathematical models*. Unpublished doctoral dissertation, Duke University, Durham, NC.
- Davis, D. G. S., & Staddon, J. E. R. (1990). Memory for reward in probabilistic choice: Markovian and non-Markovian properties. *Behaviour*, 114, 37–64.
- Davison, M., & McCarthy, D. (1988). *The matching law: A research review*. Hillsdale, NJ: Erlbaum.
- Dow, S. M., & Lea, S. E. G. (1987). Foraging in a changing environment: Simulation in the operant laboratory. In M. L. Commons, A. Kacelnik, & S. J. Shettleworth (Eds.), *Quantitative analyses of behavior VI: Foraging* (pp. 90–113). Hillsdale, NJ: Erlbaum.
- Fersen, L. von, Wynne, C. D. L., Delius, J. D., & Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17, 334–341.
- Harley, C. B. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology*, 89, 611–633.
- Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4, 267–272.
- Herrnstein, R. J., & Vaughan, W. (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits to action: The allocation of individual behavior* (pp. 143–146). San Diego, CA: Academic Press.
- Hinson, J. M., & Staddon, J. E. R. (1983). Matching, maximizing and hill climbing. *Journal of the Experimental Analysis of Behavior*, 40, 321–331.
- Honig, W. K., & Staddon, J. E. R. (Eds.). (1977). *Handbook of operant behavior*. Englewood Cliffs, NJ: Prentice Hall.
- Horner, J. M., & Staddon, J. E. R. (1987). Probabilistic choice: A simple invariance. *Behavioural Processes*, 15, 59–92.
- Hull, C. L. (1934). The concept of the habit-family hierarchy and maze learning. *Psychological Review*, 41, 33–44.
- Kacelnik, A., Krebs, J. R., & Ens, B. (1987). Foraging in a changing environment: An experiment with starlings (*Sturnus vulgaris*). In M. L. Commons, A. Kacelnik, & S. J. Shettleworth (Eds.), *Quantitative analyses of behavior VI: Foraging* (pp. 63–87). Hillsdale, NJ: Erlbaum.
- Killeen, P. R. (1981). Averaging theory. In C. M. Bradshaw, E. Szabadi, & C. F. Lowe (Eds.), *Quantification of steady state operant behaviour* (pp. 21–34). Amsterdam: Elsevier.

- Lea, S. E. G., & Dow, S. M. (1984). The integration of reinforcements over time. *Annals of the New York Academy of Sciences*, 423, 269–277.
- Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.
- Machado, A. (1992). Behavioral variability and frequency-dependent selection. *Journal of the Experimental Analysis of Behavior*, 58, 241–263.
- Mackintosh, N. J. (1974). *The psychology of animal learning*. San Diego, CA: Academic Press.
- Mazur, J. E. (1981). Optimization theory fails to predict performance of pigeons in a two-response situation. *Science*, 214, 823.
- Minsky, M. (1967). *Computation: Finite and infinite machines*. Englewood Cliffs, NJ: Prentice-Hall.
- Myerson, J., & Hale, S. (1988). Choice in transition: A comparison of melioration and the kinetic model. *Journal of the Experimental Analysis of Behavior*, 49, 291–302.
- Myerson, J., & Miezin, F. M. (1980). The kinetics of choice: An operant systems analysis. *Psychological Review*, 87, 160–174.
- Nevin, J. A. (1979). Reinforcement schedules and response strength. In M. D. Zeiler & P. Harzem (Eds.), *Reinforcement and the organization of behavior* (pp. 117–158). New York: Wiley.
- Nevin, J. A. (1988). Behavioral momentum and the partial reinforcement effect. *Psychological Bulletin*, 103, 44–56.
- Nevin, J. A., Tota, M. E., Torquato, R. D., & Shull, R. L. (1990). Alternative reinforcement increases resistance to change: Pavlovian or operant contingencies? *Journal of the Experimental Analysis of Behavior*, 53, 359–379.
- Rachlin, H. (1991). *Introduction to modern behaviorism* (3rd ed.). New York: Freeman.
- Staddon, J. E. R. (1973). On the notion of cause, with applications to behaviorism. *Behaviorism*, 1, 25–63.
- Staddon, J. E. R., & Frank, J. (1974). Mechanisms of discrimination reversal. *Animal Behaviour*, 22, 802–828.
- Staddon, J. E. R., & Hinson, J. M. (1983). Optimization: A result or a mechanism? *Science*, 221, 976–977.
- Steirn, J. N., & Zentall, T. R. (1991, November). *Transitive inference in pigeons*. Paper presented at the meeting of the Psychonomic Society, San Francisco.
- Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. Unpublished doctoral dissertation, Department of Computer and Information Science, University of Massachusetts, Amherst.
- Todorov, J. C., Castro, J. M. O., Hanna, E. S., de Sa, M. C., & Barreto, M. (1983). Choice, experience, and the generalized matching law. *Journal of the Experimental Analysis of Behavior*, 40, 99–111.
- Williams, B. A. (1976). Short-term retention of response outcome as a determinant of serial reversal learning. *Learning and Motivation*, 7, 418–430.
- Williams, B. A. (1988). Reinforcement, choice, and response strength. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., pp. 167–244). New York: Wiley.
- Williams, B. A., & Royalty, P. (1989). A test of the melioration theory of matching. *Journal of Experimental Psychology: Animal Behavior Processes*, 15, 99–113.
- Wynne, C. D. L., Fersen, L. von, & Staddon, J. E. R. (1992). Pigeons' inferences are transitive and the outcome of elementary conditioning principles: A response. *Journal of Experimental Psychology: Animal Behavior Processes*, 18, 313–315.

Appendix A

Conditions for a Performance Model

If function g is many-one, the condition that must be satisfied for a performance model is just that the composition $g(f(g^{-1}))$ is unique, which is possible for some pairs of functions f and g even if function g is many-one. For example, suppose that the model definition is

$$V_i(t+1) = kV_i(t)[h\alpha + \beta] + (1-k)V_i(t) = K_i(t)V_i(t),$$

$$\alpha > 1, 0 < \beta < 1, \quad (A1)$$

where V is the model state variable and h and k are dummy variables: $h = 1$ when the response is reinforced and 0 otherwise, and $k = 1$ if the response occurs and 0 otherwise. $K_i(t) = \{k[h\alpha + \beta] + (1-k)\}$ is thus a lumped variable summarizing the conditions for response i on iteration t . The response rule is

$$B_i(t) = V_i(t)/\Sigma V(t), \quad (A2)$$

where B_i is the probability (say) of behavior i and $\Sigma V(t)$ is the sum of V

values on each iteration. (These two equations are a version of Luce's, 1959, beta model).

Equation A2 suggests that this is a state model, because a given value for B_i is compatible with many possible values for V_i . However, in fact, if we divide both sides of Equation A1 by $\Sigma V(t+1)$, we can rewrite it solely in terms of B :

$$V_i(t+1)/\Sigma V(t+1) = B_i(t+1) = K_i(t)V_i(t)/\Sigma V(t+1). \quad (A3)$$

For just two responses, i and j , $\Sigma V(t+1) = K_i(t)V_i(t) + K_j(t)V_j(t)$. Substituting in Equation A3 and dividing top and bottom by $V_j(t)$ yields a function solely in the ratio $V_i(t)/V_j(t)$, which can be replaced by $[1 - B_i(t)]/B_i$ from Equation A2. Thus, Equations A1 and A2 can be reduced to a single equation in which $B_i(t+1) = G[B_i(t), K_i(t)]$, which is a performance model.

Note that if Equation A1 were of the form $V(t+1) = K_i(t)V_i(t) + \alpha$, where α is not a function of V , this elimination of V would be impossible.

Appendix B

Properties of INT-R

In the INT-R model, the state of the animal is represented by a vector V of response values with components V_L and V_R for L and R responses, respectively. If a left response occurs at iteration t , its V value changes in a linear way, thus

$$V_L(t+1) = aV_L(t) + (1-a)X(t), \quad 0 < a < 1, \quad (B1)$$

where $X(t)$ is reinforcement magnitude at iteration t (0 or 1 in our simulations) and a is a parameter that represents the persistence of effects.

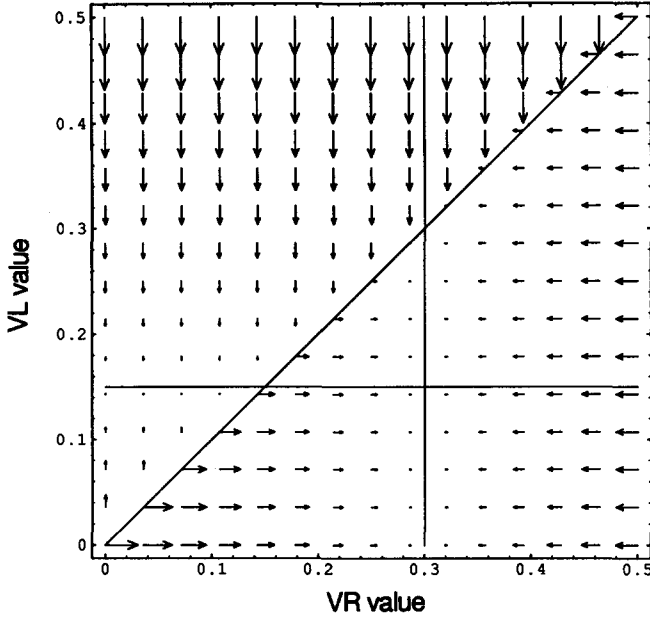


Figure B1. Phase space for the response-by-response integrator model in the two-arm-bandit situation (concurrent variable ratio-variable ratio). (The diagonal line $V_L = V_R$ represents the winner-take-all response rule: If $V_L > V_R$, a left response occurs; if $V_R > V_L$, a right response occurs. The vertical [$V_R = p$] and horizontal [$V_L = q$] lines represent the payoff probabilities for R and L responses, respectively. The arrows show the expected direction and magnitude of change in the V value after each iteration.)

The V value for the right response remains unchanged. The same process operates if a right response occurs at iteration t . The decision rule is that the response with the highest V value occurs (winner-take-all rule).

Concurrent Variable Ratio-Variable Ratio (Conc VR-VR; Two-Arm Bandit)

Let the constants p and q stand for the reward probabilities for R and L responses, respectively. The predictions of the model can be derived from its state space shown in Figure B1. The winner-take-all rule is represented by the diagonal line ($V_L = V_R$) that divides the space into two domains: above the line $V_L > V_R$, and an L response occurs; below the line $V_R > V_L$, and an R response occurs. Initially, the state of the model is represented by a point with coordinates $V_L(0)$ and $V_R(0)$. After each iteration, the point will jump to a new location whose coordinates are determined by the outcome of the iteration (see Equation B1). Given the V values at the beginning of iteration t , the arrows in the figure point to the expected location of the point at the beginning of iteration $t + 1$. Thus, when $V_L(t) > V_R(t)$, V_R remains unchanged and, on average, $V_L(t + 1)$ will equal

$$E[V_L(t + 1) | V_L(t)] = aV_L(t) + (1 - a)q, \quad (B2)$$

and, similarly, when $V_R(t) > V_L(t)$, V_L does not change and the expected value of $V_R(t + 1)$ is given by Equation B2, with q replaced by p .

From the direction field shown in Figure B1, it is clear that V_L and V_R will move in the direction of the horizontal ($V_L = q$) and vertical ($V_R = p$) lines, respectively.

While only one response is occurring, say L, Equation B1 is applied on every iteration. This stochastic difference equation has the solution

$$V_L(t) = a^t V_L(0) + (1 - a) \sum_{i=0}^{t-1} a^{t-1-i} X(i), \quad t \leq 1, \quad (B3)$$

where t is reset to 0 after each switching response. The expectation and the variance of $V_L(t)$ will equal, respectively,

$$E[V_L(t)] = a^t V_L(0) + (1 - a^t)q \quad (B4)$$

and

$$\text{Var}[V_L(t)] = q(1 - q)(1 - a^{2t-2})(1 - a)/(1 + a). \quad (B5)$$

If no switching from L to R occurs for several iterations, then t gets larger, $E(V_L)$ tends to q , and $\text{Var}(V_L)$ tends to $(1 - a)q(1 - q)/(1 + a)$. The exact distribution of the V_L values is not known. Consequently, we could not determine the probability of the event $V_L < V_R$ (for arbitrary V_R) that would allow us to predict the number of L responses before a switching occurs. However, when a is close to 1 and q is not too small, the movement along the vertical line is close to a Brownian motion; the distribution of V_L approaches the normal distribution with mean q and variance $(1 - a)q(1 - q)/(1 + a)$. The number of L responses before a switching will therefore depend on the parameters a and q and on the value of V_R .

The above limitations notwithstanding, some qualitative predictions are possible when the number of iterations is large. The critical properties of the model are shown in Figure B2. As before, the middle line represents the switching line $V_L = V_R$; the upper and lower lines represent the lines $V_L = V_R/a$ and $V_L = aV_R$, respectively, two lines that define the zone where a switching response necessarily takes place. The initial conditions, $V_L(0)$ and $V_R(0)$, are represented by number 0. Because $V_R(0) > V_L(0)$, a right response occurs first. This response is not reinforced, V_R decreases, and the coordinate point moves to the left along the horizontal line. Eventually the point crosses the switching line (square with number 1). A nonreinforced L response occurs, V_L gets below V_R , and the point crosses the switching line at 2. Then, a right, reinforced response occurs and the point moves now to the right along the second horizontal line and reaches 3. After a run of nonreinforced R responses, the point crosses the line again (at 4); a left, reinforced response is emitted, V_L increases to 5, and L responses occur until the point reaches 6. Four aspects are noteworthy in this sample path behavior of INT-R: (a) The successive switching points (1, 2, 4, 6, and 7) approach the origin; (b) each successive reinforcer for a switching response has a larger effect because ΔV is proportional to $1 - V$

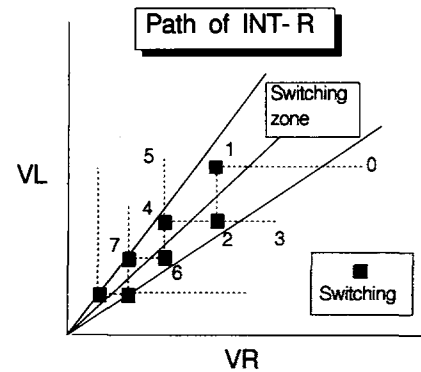


Figure B2. "Ratchet" property of the response-by-response integrator model (INT-R): sample path behavior. (The diagonal line $V_L = V_R$ represents the winner-take-all response rule. The upper and lower lines, $V_L = V_R/a$ and $V_L = aV_R$, respectively, define the switching zone. The squares represent switching points. The path starts at 0 and proceeds to 1, 2, 3, . . . Successive switching points [i.e., 1, 2, 4, 6, . . .] approach the origin. See the text for further details.)

and, at each successive switching, the V values are smaller and smaller; (c) for similar reasons, as the origin is approached, extinction has smaller effects because, in this case, ΔV is proportional to V ; (d) as a consequence of (a), (b), and (c), switching becomes less and less frequent, and more and more time is spent moving along the horizontal or vertical lines. On average, then, the length of the runs of right and left responses increases with iterations.

In a conc VR-VR, an increasing proportion of time will be spent on the majority side, R in our example. This occurs because, as Figure B1 shows, the vector field near the origin is stronger for R than for L responses ($p > q$). Although the probability of switching from R to L is never zero, the proportion of R response in any sample will get arbitrarily close to 1.

Conc VI-VI

This schedule differs from the two-arm bandit in that reinforcers are set up independently for each response alternative and, once a reinforcer is set up, it remains available until collected. As before, let p and q be the probabilities of setting up a reinforcer on each iteration for the

right and left sides, respectively. From the properties of the schedule it follows that as the run length increases, the probability of getting a reward for a (rare) changeover also increases. When compared with the preceding VR-VR, the switching point moves toward the origin with slower speed.

Concerning matching, one of two possibilities holds, depending on how large is the window over which one averages:

1. If the averaging period is large enough to include both types of responses, then undermatching is typically obtained; that is, the response ratio $x/(x+y)$ is less extreme than the reward ratio $R(x)/[R(x)+R(y)]$ (x = right key): given that $p > q$,

$$x/(x+y) \approx [R(x)/p]/[R(x)/p + R(y)/q] \quad (B6)$$

$$\approx qR(x)/[qR(x) + pR(y)] \quad (B7)$$

$$< R(x)/[R(x) + R(y)] \quad (B8)$$

(if $p < q$, then the inequality is reversed).

2. When the averaging window includes only one response, then degenerate matching results.

Appendix C

Initial Conditions for the CE Model

When V values for the two choices are approximately equal, the initial conditions implied by the CE model can be derived analytically. For a symmetrical situation, we can assume that the initial conditions for each choice are the same, so that only two numbers, initial responses, N_0 , and initial reinforcements, R_0 , need to be estimated.

A simple graphical derivation is as follows. Given the equality of V values, it follows from Equation 7 that $(R_L + R_0)/(N_L + N_0) = (R_R + R_0)/(N_R + N_0)$ at all points in training, where R_L and so forth are cumulative quantities. If we have the correct values for R_0 and N_0 , therefore, a response-by-response or session-by-session plot of cumulative $(R_L + R_0)/(N_L + N_0)$ versus $(R_R + R_0)/(N_R + N_0)$ should be a straight line with

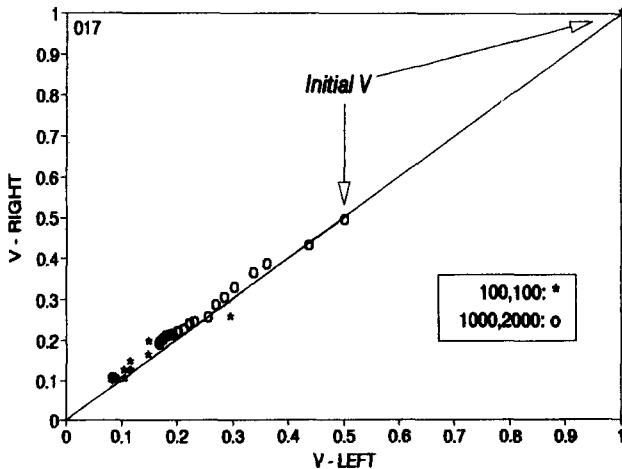


Figure C1. Estimating initial conditions for the cumulative-effects model from the data in Figure 8 (first 22 conditions). (Each point is a plot of $[R_L + R_0]/[N_L + N_0]$ versus $[R_R + R_0]/[N_R + N_0]$, with the R_0 and N_0 values shown. The points corresponding to the initial V values [100,100 or 1000,2000] are shown and later points are lower on both axes.)

unit slope. The best values for R_0 and N_0 are those that bring about the closest fit between data and the diagonal.

It turns out that as long as R_0 and N_0 are greater than about 100, the plot of $(R_L + R_0)/(N_L + N_0)$ versus $(R_R + R_0)/(N_R + N_0)$ is a pretty good fit to the diagonal. If R_0 and N_0 are very large (i.e., $R_0 \gg R_i$ and $N_0 \gg N_i$), equality between $(R_L + R_0)/(N_L + N_0)$ and $(R_R + R_0)/(N_R + N_0)$ is forced, of course, but at the cost of movement along the diagonal. However, values in the hundreds are on the same order as typical values of R_i and N_i , and do not force equality. Examples of two sets of initial conditions, 100,100 and 1000,2000, are shown in Figure C1. Each point is a session-by-session plot of $(R_L + R_0)/(N_L + N_0)$ versus $(R_R + R_0)/(N_R + N_0)$ with the R_0 and N_0 values shown (o: 1000,2000; *: 100,100) and R_i and N_i from the data in Figure 7. In both cases, the fit to the diagonal is good.

Table C1 summarizes the results of several simulations with a variety of R_0 and N_0 values. The right-most column shows the R^2 value of a regression line fitted to points like those in Figure C1, with the initial conditions shown in the table. As one can see, when R_0 and N_0 are less than 100, the fit is poor; above that value, the fit is increasingly good. Above 100,100, the slope is close to unity. For the simulation in Figure 13, we picked from the higher values in Table C1 the pair of R_0 and N_0 values that gave the best qualitative match to the data in Figure 7.

There is an algebraic method for determining the initial conditions of the CE model from a response-by-response or session-by-session series of cumulated response and reinforcement totals. From Equation 7 in the text, $V_i(t+1) = [R_i(t) + R_i(0)]/[N_i(t) + N_i(0)]$. If V values at time t are approximately equal, then for choice between symmetrical left and right responses at time t we may assume

$$[R_L + R_0]/[N_L + N_0] = [R_R + R_0]/[N_R + N_0], \quad (C1)$$

where N_0 and R_0 are initial conditions and R_i and N_i are the cumulated totals. Multiplying out and rearranging

$$(R_R N_L - R_L N_R)/(R_L - R_R) = R_0(N_R - N_L)/(R_L - R_R) + N_0, \quad (C2)$$

which may be rewritten as $Y = R_0 X + N_0$, where $Y = (R_R N_L - R_L N_R)/(R_L - R_R)$ and $X = (N_R - N_L)/(R_L - R_R)$. A response-by-response or

Table C1
Goodness of Fit of the Cumulative-Effects Model Simulation
Results and the Data in Figure 7 When Different
Initial Conditions (R_0 , N_0) Are Used

Simulation	R_0	N_0	slope	R^2
1	50	50	0.41	0.63
2	100	100	0.80	0.87
3	1000	1000	1.11	0.99
4	1000	2000	1.07	0.98
5	2000	2000	1.09	0.99
6	2000	4000	1.06	0.99
7	4000	4000	1.07	0.99
8	5000	10000	1.04	0.99
9	10000	10000	1.04	0.99

session-by-session plot of Y versus X should therefore be a straight line with slope equal to R_0 and intercept equal to N_0 .

There will be some points on such a graph for which $R_L = R_R$, and these will obviously have to be excluded. In attempting to derive initial conditions from the data in Figure 7 using this method, we have found that the slopes of the linear fits are rather steep (corresponding to R_0 and N_0 values in the thousands).

Because neither of these methods is very sensitive, and because the assumption of equal V values is sometimes violated, we chose initial values more according to the overall fit of the model to data, rather than according to estimation.

Received February 20, 1992

Revision received October 18, 1992

Accepted October 21, 1992 ■



AMERICAN PSYCHOLOGICAL ASSOCIATION

SUBSCRIPTION CLAIMS INFORMATION

Today's Date: _____

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do NOT duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION _____

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL) _____

ADDRESS _____

DATE YOUR ORDER WAS MAILED (OR PHONED) _____

CITY _____

STATE/COUNTRY _____

ZIP _____

____ PREPAID ____ CHECK ____ CHARGE

CHECK/CARD CLEARED DATE: _____

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

YOUR NAME AND PHONE NUMBER _____

ISSUES: ____ MISSING ____ DAMAGED

TITLE _____

VOLUME OR YEAR _____

NUMBER OR MONTH _____

Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4-6 weeks.

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: _____

DATE OF ACTION: _____

ACTION TAKEN: _____

INV. NO. & DATE: _____

STAFF NAME: _____

LABEL NO. & DATE: _____

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.