Journal of Experimental Psychology:
Animal Behavior Processes
2001, Vol. 27, No. 4, 354–372

Copyright 2001 by the American Psychological Association, Inc.
0097-7403/01/$5.00 DOI: 10.1037//0097-7403.27.4.354

The Rat Approximates an Ideal Detector of Changes in Rates of Reward: Implications for the Law of Effect

C. R. Gallistel and Terence A. Mark
University of California, Los Angeles

Adam Philip King
Fairfield University

P. E. Latham
University of California, Los Angeles

Rats responded on 2 levers delivering brain stimulation reward on concurrent variable interval schedules. Following many successive sessions with unchanging relative rates of reward, subjects adjusted to an eventual change slowly and showed spontaneous reversions at the beginning of subsequent sessions. When changes in rates of reward occurred between and within every session, subjects adjusted to them about as rapidly as they could in principle do so, as shown by comparison to a Bayesian model of an ideal detector. This and other features of the adjustments to frequent changes imply that the behavioral effect of reinforcement depends on the subject's perception of incomes and changes in incomes rather than on the strengthening and weakening of behaviors in accord with their past effects or expected results. Models for the process by which perceived incomes determine stay durations and for the process that detects changes in rates are developed.

When subjects of many different species choose between response options that are rewarded on concurrent variable interval schedules, the ratio of the amounts of time they invest in the options approximates the ratio of the incomes they realize from them, where income is defined as the amount of reward per unit of session time (Davison & McCarthy, 1988; Herrnstein, 1961, 1991; Herrnstein & Prelec, 1991). This is called matching behavior. It is at least approximately optimal in that there is no other response pattern that will substantially increase the overall income (Baum, 1981; Heyman & Luce, 1979). The question is, how is this income-maximizing pattern of behavior arrived at? Is it shaped by the selective strengthening and weakening of responses by their consequences? Or is matching elicited by the perceived ratio of the incomes, as Heyman (1982) has argued?

A variable interval (VI) schedule is a random rate process with exponentially distributed delays of reward. The scheduling of the next reward is independent of the time that has elapsed since the last reward was harvested. Importantly, once a reward is scheduled, it remains available until the subject again tries that option (depresses the lever served by that schedule) and thereby harvests the reward. Thus, the longer the subject has gone without pressing one of the levers, the more certain it is that a reward is set up there for immediate harvest. When two levers are concurrently available, each producing rewards through independently operating VI schedules, a subject can obtain something approaching the maximum possible combined income if it repeatedly tries both levers at intervals shorter than the expected delays (Heyman, 1982). Even when the levers pay off at very different rates, the subject gets more income by moving back and forth between them than it would get by devoting all of its time to the lever that pays off more frequently. Thus, matching behavior in the face of these contingencies is rational in the economists' sense of the term.

Instrumentally conditioned or operant behavior is by definition the result of a hill climbing process involving feedback from the consequences of behavior onto the mapping from a perceived stimulus situation to the behavior produced by that perception. In the given situation, different behaviors are tried. Those that produce reward are strengthened; those that produce no reward are weakened. Situation-specific response strengthening and weakening continues until the subject arrives at a pattern of behavior that maximizes its return, the amount of reward per response or per unit of time invested (Herrnstein & Vaughan, 1980). Elicited (or unconditioned) behavior, by contrast, is generated by a purely feedforward process; the conditioned response is elicited by a given perception. In classical or Pavlonian conditioning, for example, the perception of a temporal contingency between a neutral stimulus (CS) and a reinforcer elicits conditioned responding to the CS, whether those responses have reinforcing consequences or not (Brown & Jenkins, 1968; Gamzu & Williams, 1971; Williams & Williams, 1969).

C. R. Gallistel and Terence A. Mark, Department of Psychology, University of California, Los Angeles; Adam Philip King, Department of Computer Science, Fairfield University; P. E. Latham, Department of Neurobiology, University of California, Los Angeles.

This work was supported by National Science Foundation Grant IBN-93062383.

Data accessible at http://ruccs.rutgers.edu/rat_data.

Correspondence concerning this article should be addressed to C. R. Gallistel, who is now at Rutgers Center for Cognitive Science, Psychology Building, Busch Campus, 152 Frelinghuysen Road, Piscataway, New Jersey 08854-8020. Electronic mail may be sent to galliste@ruccs.rutgers.edu.

Purely feed-forward processes can adjust to changes in the stimulus situation more rapidly than feedback processes because the completion of the adjustment does not require the assessment of the consequences of intermediate adjustments. Thus, measuring the rapidity of the adjustments to changes in the relative rates of reward may reveal the locus of reinforcement's effect. Does it act either on the strength of a stimulus–response connection or on the expected values of responses, as models involving selection by consequences posit? Or does it affect only perceived incomes, with the observed behaviors being fixed consequences of those perceptions, as purely feed-forward models posit?

In experiments with rats pressing two levers for brain-stimulation rewards delivered on concurrent variable interval schedules, we made signaled and unsignaled step changes in the scheduled rates of reward and measured how rapidly subjects adjusted the expected durations of their stays. We found that when these changes occur frequently, the subjects' adjustments are about as rapid as they could in principle be.

In previous work (Mark & Gallistel, 1994), we found that rats adjusted to a signaled change in the relative rates of brain-stimulation reward extremely rapidly—within one or two interreward intervals on the learner schedule. Dreyfus (1991) reported comparable findings using pigeons responding for food reward. His changes occurred after a fixed amount of session time had elapsed but were otherwise unsignaled. In both cases, however, the subjects experienced frequent changes in the relative rates of reward. In Mazur's (1995) experiments, by contrast, pigeons experienced a prolonged multisession stability in the relative rates of reward before an unsignaled change. The transitions observed in Mazur's experiments took considerably longer than those observed by Mark and Gallistel (1994) and Dreyfus (1991).

In the experiments we now report, our subjects first experienced a multisession phase with unchanging relative rates of reward, then an unsignaled transition, then another prolonged phase of constancy at the new relative rates, then phases in which the relative rates and overall rates changed frequently, then, finally, another prolonged phase of constancy followed by a final transition. This design allows us to contrast the transitions seen after prolonged stability with the transitions seen when the relative rates of reward change frequently. It also allows us to determine how many changes in the relative rates of reward a rat must experience before its transitional behavior changes from the slow transitions that follow prolonged stability to the rapid transitions seen when the rates of reward change frequently. Finally, it allows us to determine whether slow transitions are due to lack of prior experience with changes in rates of reward or to prolonged stability preceding a change.

Method

Subjects

The subjects were 6 white male Sprague-Dawley rats (bred at the University of California, Los Angeles), implanted with monopolar stimulating electrodes in the posterior part of the lateral hypothalamus. The electrodes were made from No. 00 stainless steel insect pins insulated with Formvar to within 0.5 mm of their tip. There was an indifferent electrode on the skull surface. The placement of the stimulating electrodes was verified by standard histological procedures at the conclusion of the experiments. The subjects were 100–120 days old and weighed between 290 and 400 grams when they were implanted. These subjects were selected from a larger pool of similarly implanted subjects because they learned to press a lever for brain stimulation reward (at the parameters indicated below) during a 30 min screening session conducted 1 week after the electrodes were implanted. They entered the initial phase of the experiment after two 30 min sessions of continual reinforcement in a single-lever screening box.

Apparatus

The subjects responded for brain stimulation reward in Plexiglas boxes measuring 26 $cm^2 \times 44$ cm high. The floors were covered with hardware cloth. In the center of one wall, a Plexiglas protrusion created two alcoves, 11.5 cm wide × 11 cm deep. In the center of the wall at the back of each alcove, there was a retractable lever (BRS/LVE Model RRL-015) located 5.5 cm above the floor. Placing the levers in alcoves made it impossible for a subject to switch between levers in less than about 1.5 s. Thus, there was no need for a changeover delay, which is often added to concurrent variable interval schedules to prevent subjects from alternating very rapidly between levers or keys.

The experiments were controlled by PC/XT type microcomputers, which controlled custom-designed constant-current stimulators. The stimulators shunted the stimulating electrode to the indifferent electrode between pulses, thereby preventing the monophasic cathodal pulses from polarizing the electrode–tissue interface. The computers specified all parameters of the stimulation, scheduled the rewards, and recorded the data. The data were logged in the form of a text file, with each line in the file representing the occurrence of a specified event and the elapsed session time (in milliseconds) at which it occurred. The events recorded were lever1 down, lever1 up, lever2 down, lever2 up, reward1 armed (set up by the schedule), reward2 armed, reward1 delivered, and reward2 delivered. For technical reasons, the session timer did not run during the half second when a reward was delivered. Thus, reward deliveries appear in the raw data as point events, events with no duration.

Procedure

The rewards were 0.5-s trains of 0.1-ms cathodal pulses, delivered at a frequency of 126 pulses per second and an amplitude of 400 μA. VI schedules of reinforcement were generated using independent constant probability geometric approximations to an exponential distribution. Beginning immediately after a reward was collected from a lever, the computer flipped an electronic coin once each second to determine whether to set up the next reward on that lever. The probability, p, of this coin coming up "heads" determined the expected delay ($1/p$ s) to the next available reward. The reward was delivered immediately if the rat was holding the lever down at the moment it was set up. If not, the reward was delivered on the next depression of the lever.

The experiment was run in approximately 130 daily sessions lasting 2 hr each. The first phase of the experiment was 33.5 sessions long (36.5 sessions in Subject Rx). The relative rates of reward on the two sides remained constant throughout this phase. The second phase began half way through the 34th session (37th for Rx), when there was an unsignaled change in the relative rates of reward. The new relative rates of reward remained in force throughout the second half of the session and for 20 sessions thereafter. The third and fourth phases, each lasting 20 sessions, began with the session following the end of the second phase. In these two phases, subjects experienced a change in the rates of reward between the end of each session and the beginning of the next. Also, within each session, they experienced an unsignaled change at a point selected randomly with uniform probability within the middle 80 min of the session. The sequence of conditions for each subject is given in Table 1.

Phases three and four differed with regard to whether it was the relative rates or the overall rates that changed. In one phase, called the *relative*

Table 1
Sequence of Conditions by Subject

	Phase					
	1	2	3	4	5	6
No. of 2-hr sessions	33.5	20.5	20	20	33.5	20.5
	Programmed rates of reward					
Subject						
A	1/1, 9.4	1/4, 9.4	varied, 9.4	1/1, varied	1/1, 9.4	1/4, 9.4
B	1/1, 9.4	4/1, 9.4	1/1, varied	varied, 9.4	1/1, 9.4	4/1, 9.4
D	1/4, 9.4	4/1, 9.4	varied, 9.4	1/1, varied	1/4, 9.4	4/1, 9.4
E	1/4, 9.4	4/1, 9.4	1/1, varied	varied, 9.4	1/4, 9.4	4/1, 9.4
K	1/1, 9.4	1/4, 9.4	varied, 9.4	1/1, varied	1/1, 9.4	1/4, 9.4
Rx	1/1, 9.4	4/1, 9.4	varied, 9.4	1/1, varied	1/1, 9.4	4/1, 9.4

Note. The programmed rates of reward are presented as Side 1/Side 2 and total rewards per minute.

condition, the sum of the programmed rates—hence the programmed overall rate—was held constant (at 9.4 rewards/min); only the relative rates changed. In the other phase, called the *overall condition,* the relative rates of reward were fixed at 1:1, and the sum of the rates varied between and within each session. (The sum varied between the following values: 2.1, 6, 9.4, and 18 rewards per minute.)

In the fifth phase, the subjects ran further 33.5 sessions with unvarying relative rates of reward. At the beginning of the sixth phase, there was an unsignaled midsession change in the relative rates of reward, which remained in force for several more sessions. The purpose of the fifth and sixth phases was to determine whether a renewed period of prolonged stability would lead to slow transitions in subjects that had previously made very rapid transitions.

When the relative rates of reward varied, five pairs of schedules were used (VI 7.1-s/VI 62.5-s, VI 8.55-s/VI 25.64-s, VI 12.82-s/VI 12.82-s, VI 25.64-s/VI 8.55-s, and VI 62.5-s/VI 7.1 s). The corresponding ratios of rates of reward are 9/1, 3/1, 1/1, 1/3, and 1/9. The sum of the rates in each pair (the sum of the reciprocals of the VIs) is 9.4 rewards per minute. For some subjects, the between-session transitions were initially big (from 9/1 to 1/9) and grew progressively smaller over sessions. In these subjects, the within-session transitions were initially small (e.g., from 1/1 to 1/3) and grew progressively bigger over sessions. For other subjects, these orders were reversed. The factor by which the rate on any one side changed varied from 1.2 to 9.

The between-session transitions are signaled transitions because a change in the rates of reward reliably occurs at the beginning of each such session. The within-session transitions are unsignaled. Thus, some subjects first encountered big unsignaled transitions and small signaled transitions, whereas other subjects first encountered small signaled transitions and big unsignaled transitions. All subjects, however, eventually experienced the full range of both kinds. In the case of the signaled transitions, subjects experienced step transitions of unpredictable direction and magnitude at a predictable time (the beginning of a session). In the case of the unsignaled transitions, subjects experienced transitions of unpredictable direction and magnitude at an unpredictable time within the middle two thirds of each session.

For our time-allocation analyses, we divided session time into four mutually exclusive and exhaustive categories: stays on Side 1, travel time from Side 1 to Side 2, stays on Side 2, and travel time from Side 2 to Side 1. A stay on Side 1 lasted from the first lever down on Side 1 following a lever up on Side 2 to the last lever up on Side 1 before a lever down on Side 2—and similarly for a stay on Side 2. In this report, we focus on an analysis of the stay durations because matching behavior is defined by the relation between the expected stay durations on the two sides and the expected interreward intervals.

Results

Distributions of Stay Durations

The distributions of stay durations (dwell times) rose to a peak between 0.5 and 1 s and then tailed off in an approximately exponential manner, regardless of the relative rates of reward and of whether it was the distribution for the richer or the learner side (Figure 1). This is consistent with previous reports in the pigeon (Gibbon, 1995; Heyman, 1979, 1982).

In an exponential distribution, the fraction of the total number of visits lasting longer than a given interval decreases by a

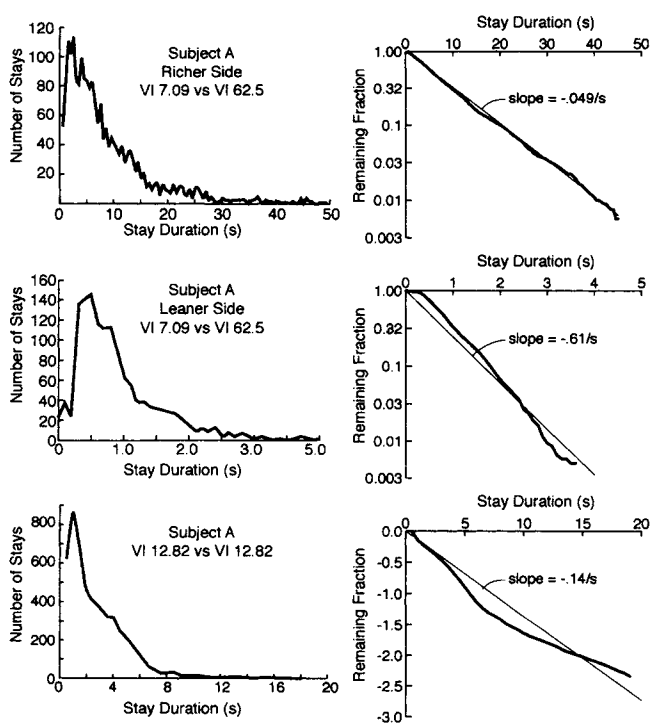

Figure 1. Representative distributions of stay durations (left side) and the corresponding log survivor plots (right side).

fixed percentage per unit increase in the given interval. Hence, the logarithm of the surviving fraction is a linearly decreasing function of time (right side of Figure 1). The slope of a log survivor plot gives the leaving rate in departures per second (after a sign change). (The slopes of the plots in Figure 1 must be multiplied by 2.3 to obtain the leaving rate because we used the common logarithm.) The leaving rate is the reciprocal of the expected stay duration. Thus, in the distribution at the top of Figure 1, there were $2.3 \times 0.049 = 0.11$ departures per second of time spent on the richer side in all those semisessions with a 9:1 ratio of programmed rates of reward. The expected duration of a stay on this side was $1/0.11 = 8.86$ s. By contrast, the expected duration on the leaner side (middle plot) was 0.71 s $= 1/(2.3 \times 0.61)$. The expected duration of a stay in those semisessions with equal programmed rates of reward (concurrent VI 12.8 s schedules) was 3.1 s $= 1/(2.3 \times 0.14)$, which is somewhat shorter than half the expected interval between successive rewards (6.4 s). When subjects sample the two sides at these rates, variations in the relative amounts of time they allocate to the two sides have little impact on the incomes they obtain (Heyman, 1982).

As noted above in the Methods section, for technical reasons, the session clock did not run while the computer was delivering a 0.5-s reward. In consequence, the raw data records occasionally have stays of 0 duration. These occurred whenever the lever was armed before the rat depressed the lever to begin its stay (i.e., there was a reward waiting to be harvested); the rat released the lever while the reward was being delivered and then left to return to the other side. These bogus 0-duration stays cause wild jumps in our statistical measure of the strength of the evidence for changes in the expected stay duration (developed below). Therefore, before applying our algorithms to the data, we corrected the raw data record by changing stays of 0 duration to stays of 0.5-s duration. This value was chosen for two reasons: (a) It was the duration of the reward, and (b) it was the most probable stay duration in the condition that produced the shortest expected stay duration (middle panel of Figure 1).

Transitions After Prolonged Stability

To visualize the time course of a transition from one pattern of time allocation to another, we need to portray the evolution of the subject's time-allocation behavior. The portrayal that has proved most illuminating is the graph of the cumulative time on Lever 1 against the cumulative time on Lever 2, hereafter called the cum-cum function. The x coordinate of a point on this graph is the cumulative duration of the stays on Side 2 up to a given moment in the session; the y coordinate is the cumulative duration of the stays on Side 1 up to that same moment. The slope at any point is the subject's time-allocation ratio at that point. When the animal matches, the slope of this function equals the ratio of the rates of reward. For example, if the rate of reward on Side 1 is four times slower than the rate on Side 2, the subject is matching when the average duration of a stay on Side 1 is four times shorter than the average duration of a stay on Side 2. In that case, the slope of the cum-cum function is 1/4—1 min spent on Side 1 for every 4 spent on Side 2 (see Figure 2).

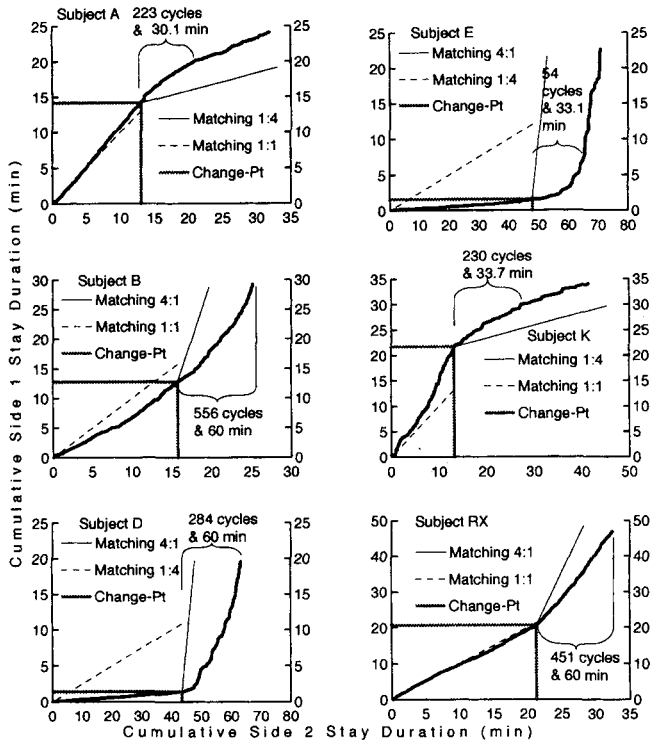

Figure 2. Cum-cum functions for the first (Phase 1–Phase 2) transition. The thin matching lines indicate the programmed ratio of reward rates. When the cum-cum curve parallels this line, the subject's time-allocation ratio matches the ratio of the programmed rates of reward. Thick gray lines show the cum-cum coordinates of the programmed change point, the total amounts of time spent on each side up to the moment at which the programmed rates of reward changed.

Ideally, the matching lines that we show on cum-cum figures for comparison purposes would express the ongoing ratio of the obtained rates of reward rather than the ratio of the programmed rates. This would enable us to compare the slope of the cum-cum function to the slope of the rate ratios actually experienced by the subject. The evolving rate ratio, however, cannot be displayed on the same axes as the cum-cum function because it is the ratio of the number of rewards on Side 1 to the number of rewards on Side 2 ($n_1/t/n_2/t = n_1/n_2$, where t is session time). The ratio of the obtained rates of reward is generally close to the ratio of programmed rates of reward. Nonetheless, it should be born in mind that in looking at the evolution of a subject's time allocation by means of a cum-cum function, we compared the slope of that function against the ratio of the programmed rates of reward, not the obtained rates, so some of portion of an apparent discrepancy in slopes (apparent failure to match) is due to differences between the programmed relative rate of reward and the actually obtained relative rate.

The point at which the cum-cum functions in Figure 2 became approximately parallel to the postchange matching line was estimated by eye, and the interval from that point back to the programmed change point was measured, both in session minutes and visit cycles (number of visits to each side). The intervals thus determined were substantial. The subjects took from 30 to 60 min

and 54 to 550 visit cycles to adjust completely or nearly completely to the unexpected change in the rates of reward.[1]

Although the subjects took a substantial interval to complete their adjustment, the adjustments began soon after the programmed change. To substantiate this claim, we needed a running measure of the evidence for changes in rates of reward and of the evidence for changes in the rat's behavior. In the Appendix, we develop a Bayesian formula for calculating the log of the odds (the logit) that there has been a change in a random rate process. A logit of 2 corresponds to odds of 100 to 1 against the null hypothesis (no change in rate), that is, to a p value of .01 in a conventional test for significance; a logit of 3 corresponds to odds of 1,000 to 1 ($p <$ .001); and so on. In Figure 3, we plot these logits against session minutes for the period beginning 5 min before the programmed change and extending until the evidence for a change exceeds a p value of .0001 (logit = 4).

When there is a big change in reward rate on one side and a much smaller change on the other, strong evidence for these changes appears first on the side where the change is large and only later on the other side. Similarly, evidence of a change in the rat's expected stay duration often appears sooner on one side than on the other. To show the lag between the evidence for a change in reward rates and the evidence for a change in stay durations, we plotted the earlier rising reward curve and the earlier rising stay duration curve. In other words, we plotted the lag between the first evidence for a change in rates of reward and the first evidence of a change in expected stay durations (Figure 3).

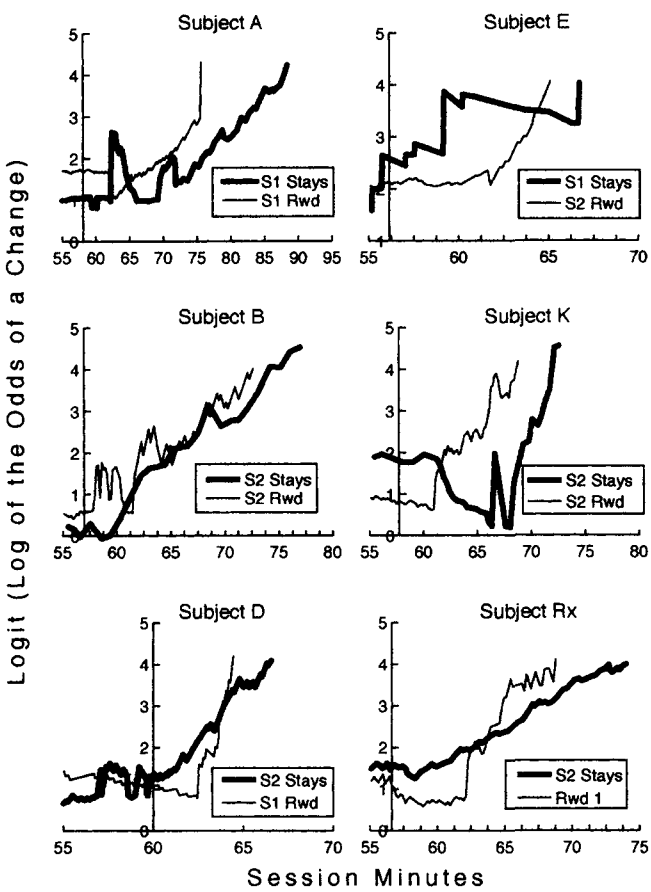

Figure 3. The emergence of the earliest evidence for a change in reward (Rwd) rates and the earliest evidence for a change in stay durations as a function of session time at the first transition. The ordinate intersects the abscissa at the moment of the programmed change in reward rates. In cases where there was already strong evidence for a change in stay durations at the time of the programmed change in rates of reward, the curves for the stay duration series have been lowered to bring their commencements close to those of the other curves (Subjects B and K).

Figure 3 shows that the latencies between the appearance of evidence for a change in rates of reward and the appearance of evidence for changes in the subject's stay duration is short—on the order of minutes or less. Thus, subjects detected the change almost as soon as there was evidence of it and began their adjustment to it at that time, but the time course of the adjustment was prolonged.

Figure 4 shows the cum-cum functions for the final transition. When the subjects experienced this change in the relative rates of reward, they all had extensive experience with such changes, but this one was preceded by a long period of stability, like the first transition. The cum-cum functions in Figure 4 (final transition) are strikingly similar to the cum-cum functions in Figure 2 (first transition). Thus, the amount of prior experience with changes in rates of reward does not determine how rapidly a subject completes its adjustment to a change in the relative rates of reward. What matters is the frequency with which such changes have been encountered recently. When changes have been infrequent, the subject takes a long time to complete its adjustment. This is as true when subjects have experienced many changes in rates of reward as when subjects have no previous experience of such changes. By contrast, when changes have recently been frequent, the subject completes its adjustment extremely rapidly (see below).

Reversions to the Status Quo Ante

Mazur (1995, 1996) reported that when pigeons experienced an unsignaled midsession change in concurrent VI schedules, which remained in force through subsequent sessions, their time-allocation behavior reverted to the status quo ante at the beginning of the next few sessions. We saw the same thing in our rat subjects following the first and final transitions (Figure 5). This spontaneous recovery of earlier time-allocation behavior explains the informally reported experience of many investigators that it takes several sessions for matching behavior to stabilize following a change in the relative rates of reward. As is seen below, this multisession adjustment is only seen when there has been a long period of stability prior to the change.

Abrupt and Complete Adjustments When Changes Are Frequent

In the relative condition, where there was a new relative rate of reward at the beginning of each session and an unsignaled change

[1] The subjects with prolonged exposure to a fourfold difference in reward rates during Phase 1 showed strong overmatching after the first four or five sessions; that is, they spent much more than four times as much time on the richer side, as is evident in Figure 2 and again in Figure 4. Strong overmatching has rarely been reported. We are not sure why it emerged during the prolonged stability phase of this experiment. We conjecture that it may be related to the high reward densities.

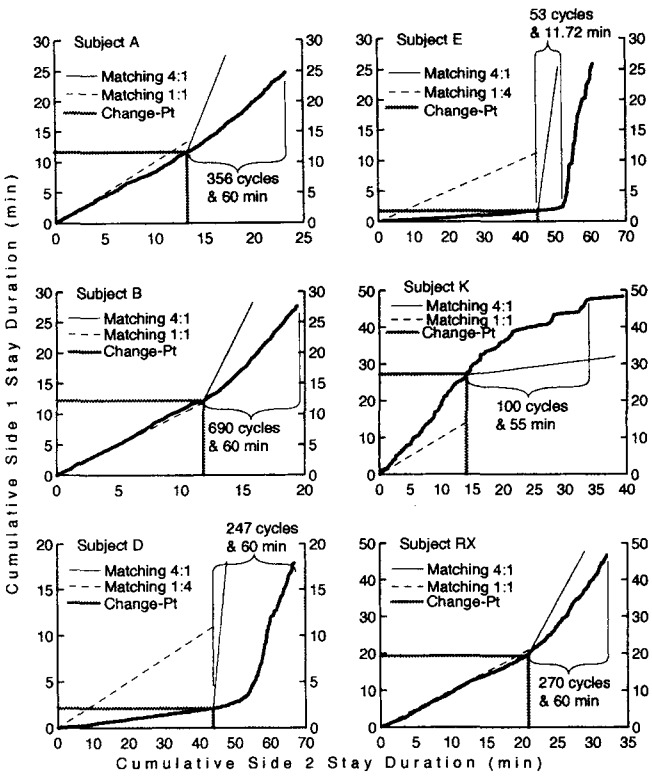

Figure 4. Cum-cum functions from the final transitions (Phase 5–Phase 6 transition). As in Figure 2, the point at which the cum-cum function became approximately parallel to the postchange matching line (thin solid lines) was estimated by eye, and the interval to achieve this adjustment was measured both in session minutes and visit cycles. Heavy gray lines give the coordinates of the cum-cum function at the moment of the programmed change in rates of reward. The dashed lines are the prechange matching lines.

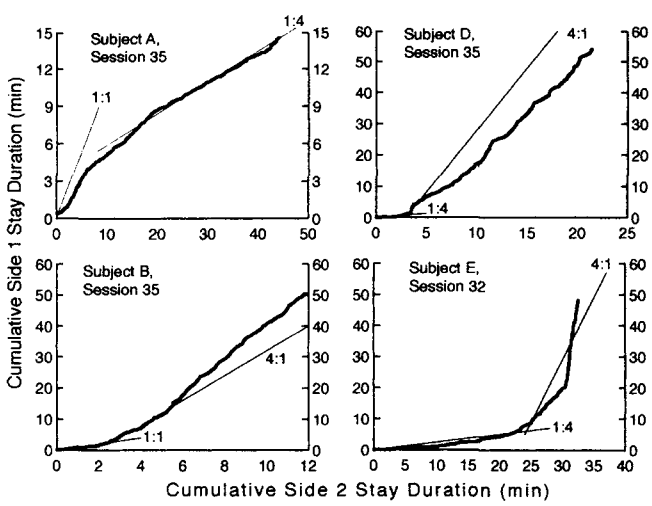

Figure 5. At the beginning of the session following the first transition, subjects' time-allocation behavior reverts to the pattern preceding the change in the relative rates of reward (representative data). On each graph, we have drawn for comparison purposes the matching line corresponding to the pretransition rate ratio and the matching line corresponding to the posttransition rate ratio.

somewhere in the middle, subjects generally showed approximate matching to each relative rate of reward, and they soon began to adjust completely within the span of a very few visit cycles. From the insets in Figure 6, which plot the adjustments visit cycle by visit cycle, it can be seen that the entire shift in behavior from the prechange time-allocation pattern to the postchange time-allocation pattern often occurred within the span of one or two visit cycles. Time-allocation ratios did not slowly approach a new steady state; they jumped from the old steady state to the new steady state.

Adjustment Latencies

It is evident from Figure 6 (see also Figure 11) that the abrupt adjustments in the rat's expected stay durations occur soon after the programmed change in the relative rates of reward. How quickly does the observed change in behavior follow the observed change in reward rates? To answer this question, we must compare an estimate of the time at which the rat reacted to an estimate of the time at which reward rates changed. Neither time can be known exactly by an observer not privy to the computer program scheduling the rewards. Their values must instead be described by probability density functions (pdfs). These pdfs, one for the change

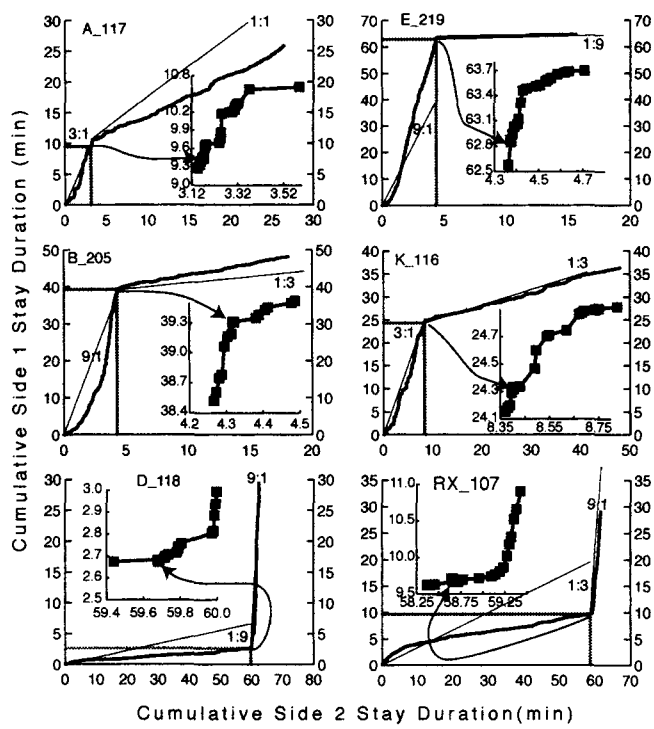

Figure 6. Sample cum-cum functions showing the abrupt transitions generally observed in the relative condition in response to the unsignaled within-session changes in the relative rates of reward. The thin lines show the programmed relative rates of reward. The thick gray lines (and the gray squares on insets) show the coordinates of the programmed change point. The insets show the cum-cum function in the immediate vicinity of its "knee," the period during which the subject adjusted its average stay duration to the new relative rates of reward. The points in these insets represent the cum-cum function at the termination of each successive visit cycle. The labels on each panel identify the subject and session number.

in rates of reward and one for the change in leaving rates, can be computed from the sequence of interreward intervals and the sequence of stay durations, respectively. (The derivation of this computation is given in the Appendix; the relevant equation is A4). Our measure of how soon the behavior changed following changes in reward rates was the normalized distance between the reward pdf with the earlier mean, min ($\bar{t}_{r1}$, $\bar{t}_{r2}$), and the stay duration pdf with the earlier mean, min ($\bar{t}_{d1}$, $\bar{t}_{d2}$)—see Figure 7A. The normalized distance, which we denote the *lag measure,* is defined to be the t statistic, that is, the difference between the two means divided by the square root of the sum of their variances (Figure 7B).

Figure 8 shows the distribution of these lag measures for the midsession transitions in the relative condition. As one would expect, the lags are mostly positive (80%), meaning that the first evidence for a change in stay durations appeared after the first evidence for a change in rates of reward. They are also mostly small, meaning that there was little lag between the first evidence of a change in reward and the first evidence of a change in behavior.

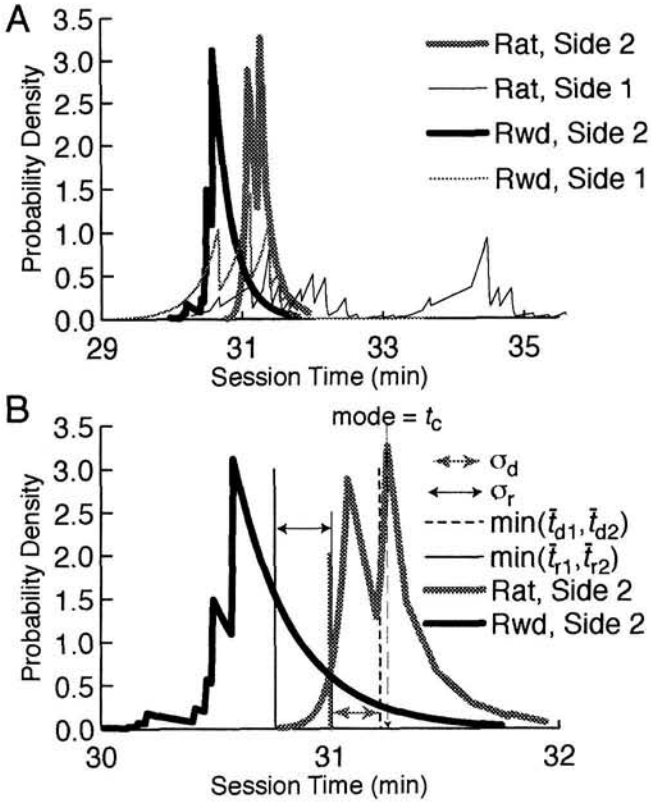

Figure 7. Panel A: Representative temporal probability density functions (pdfs), specifying the location within the session of the changes in the two rates of reward (Rwd, Side 1 and Rwd, Side 2) and the answering changes in the rat's stay durations (Rat, Side 1 and Rat, Side 2). In this example, the pdf for the reward change that first becomes evident (thick black line) is on Side 2 and so is the pdf for the change in stay durations that first appears (thick gray line). These are the two earlier pdfs. Panel B: Replotting of the two earlier pdfs from Panel A. The difference between the means of these pdfs (min[$\bar{t}_{d1}$, $\bar{t}_{d2}$] − min[$\bar{t}_{r1}$, $\bar{t}_{r2}$]) divided by the square root of the sum of their variances is a measure of how closely the first evident change in expected stay duration followed the first evident change in reward rates. The mode of the pdf for the change in the rat's stay durations is the moment, t_c, at which that change is maximally likely to have occurred.

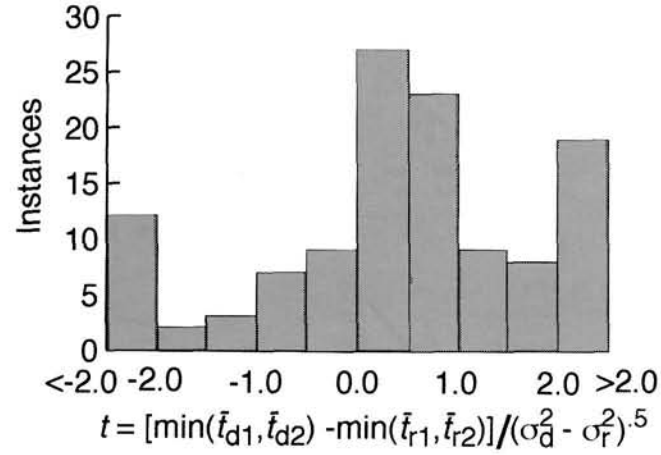

Figure 8. Histogram of the statistical distances (t statistics) by which the change in stay durations lagged the change in reward rates.

Negative lags come from cases in which the mean of the stay duration pdf was earlier than the mean of the reward pdf. These arose in one of two ways: (a) small negative lags occurred when the rat made a large change in its stay durations even though there was only weak evidence for a change in the rates of reward, and (b) large negative lags occurred when there was a spontaneous change in the rat's stay duration prior to the change in rates of reward (for examples of such changes, see Figure 13). These anticipatory changes were presumably fortuitous; they occurred less than 10% of the time.

Figure 8 shows that the lags were generally small. However, the lag is not a measure of how well the rats did relative to what was in principle possible because it is not a real-time measure: In calculating the pdfs for both the change in reward rates and the change in expected stay durations, we made use of the information from the whole session. We should ask, how well could the rat do using only information that is available in real time? As with the pdfs described above, the answer is probabilistic: The best any observer can do is compute the probability density distribution of the reward rate given the observed interreward intervals.

In the Appendix, we show how to compute such a probability density distribution—see especially Equation A12. Shown in Figure 9 are these probability density distributions; probability density (y axis) is plotted as a function of rate (x axis) at successive session times (z axis). In the left-hand panels of Figure 9, successive pdfs (successive curves in the x–y plane) specify probabilistically the reward rate on the basis of the information contained in the sequence of interreward intervals up to that point in the session. Early in the session, after only a few rewards, the reward rate is both poorly defined (a shallow broad curve) and misleading (the curves peak in the wrong place). As more rewards are obtained, however, the curves become narrow, sharply peaked and consistently located in the same region of the rate axis. At Minute 31 of the session, the programmed reward rate changed. At this point, the pdf for the reward rate became transiently shallower and broader, but it soon rose and resharpened as the new rate of reward became well defined. The broadness near the transition point reflects the uncertainty whether the rate has changed or not.

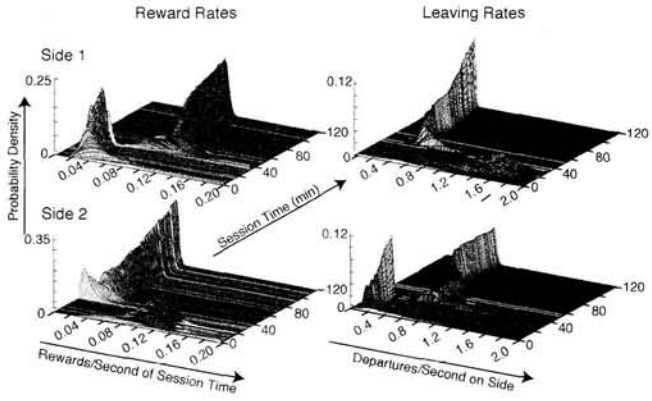

Figure 9. Left panels: Probability density (y axis) for reward rates (x axis) as a function of session time (z axis). The clock for the interreward intervals runs continuously. Right panels: Probability density for the rat's leaving rate as a function of session time. The leaving rate is the number of departures per unit of time on side; that is, a stay–duration clock runs only when the rat is on a given side. We mapped the latter functions onto session time using the session times associated with each departure.

In the right-hand panels of Figure 9, successive pdfs specify probabilistically the rat's leaving rate on a given side (the reciprocal of its expected stay duration) on the basis of the information available up to that point in the session. Notice that in the period before the change in the relative rates of reward, there were several spontaneous changes in the rat's expected stay durations. Despite this instability in its prechange behavior, there is a clear reaction to the change in the programmed rates of reward; in response to this increase in the relative rate of reward on Side 2, the rat's expected stay duration abruptly shifted to a longer time (a lower leaving rate), whereas on Side 1 it abruptly shifted to a shorter time. The abruptness of the changes in the pdfs for the rat's stay durations is comparable to the abruptness of the change in the pdfs for the reward rates, and the latter changes are known to be step changes. Thus, as already indicated (Figure 6), the rat's adjustment approximates a step adjustment.

The pdfs in Figure 9 were derived using a Bayesian approach combined with a hidden Markov model (see the Appendix for details). The basic idea behind the derivation is as follows: Given the time the reward rate changes and the rates before and after that change, one could write down the joint probability distribution for reward times. That joint distribution can be inverted, using Bayes' theorem, to derive the probability distribution of reward rates and time of change given the set of reward times. The latter distribution, however, depends on the prior for the reward rates, a prior that evolves in time. To take into account this time-dependent prior, a hidden Markov model (DeWeese & Zador, 1998) must be used: The reward rates and time of change are hidden, whereas the reward times are observed.

The derivation sketched above leads to a differential equation for two quantities: the time-dependent probability density distribution for the reward rates before the rate changed, $P_1(r, t)$, and after the rate changed, $P_0(r, t)$. The integrals of these two distributions with respect to r sum to one and the odds that a change has occurred are given by the ratio of the integrals. Initially, $P_1(r, t)$ has most of the probability whereas $P_0(r, t)$ has little. Thus, the odds that the change has already occurred are small. After the change in rate, however, the probability shifts progressively to $P_0(r, t)$. As it does, the odds that the change has occurred grow large.

We can interpret the area under $P_0(r, t)$, denoted $P_0(t)$ and defined by $P_0(t) = \int P_0(r, t)dr$, as the probability that the reward rate had changed as of any given session time. The question is, how great was this probability when the rat changed its stay durations? How strong was the evidence of change that caused the rat to change its behavior? An estimate of the time at which the rat changed its expected stay durations is the mode of the earlier pdf (see dashed arrow labeled *mode* in Figure 7B). This is the time of maximum likelihood for the behavioral change. If we denote this estimate of the time at which the rat made its decision by t_c, then $P_0(t_c)$ is the probability that the reward rates had changed prior to that time. This probability is an estimate of the rat's decision criterion. In Figure 10 we plot the distribution of these estimates. Note that in 80% of the 118 transitions used, the objective probability that the reward rates had changed was less than 0.99 at the moment when the rat decided to change its expected stay durations. This corresponds to a decision criterion of $p < .01$ in a conventional test of a null hypothesis. The data in Figure 10 imply that the rat made accurate assessments of the evidence for a change in rates of reward and used a low decision criterion. A stricter decision criterion would shift the data in Figure 10 to the right because it would lead to a delayed reaction on the part of the rat. Any insensitivity on the rat's part to the objective evidence for a change in rates of reward would also delay its decision.

The data in Figures 8 and 10 suggest that the rat approximates an ideal detector of changes in the rates of reward; it detects a change about as soon as it is objectively possible to detect it. We have already seen that the rat then adjusts to that change about as abruptly as it could. It shifts from a ratio of expected stay durations that approximately matches the old ratio of reward rates to a ratio that approximately matches the new ratio of reward rates in the span of a few visit cycles.

Rapid Emergence of Abrupt Adjustments

The shift from the prolonged transitions seen in Figures 4 and 5 to the abrupt transitions seen in Figure 6 occurred within the first

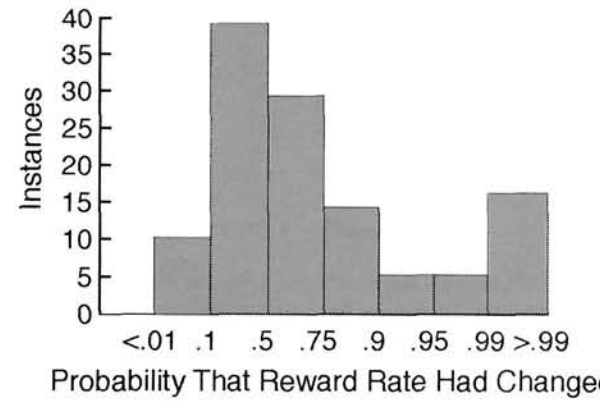

Figure 10. Histogram of the objective probabilities that the reward rates had changed at the moment at which the first change in the rat's stay durations most likely occurred.

two to four sessions, as may be seen from Figure 11, which shows the transitions for each subject in the third or fourth session after the beginning of the relative condition. At this point in the experiment, the subjects had only experienced four or five unsignaled midsession changes in the relative rates of reward. Thus, a limited experience with frequently occurring changes in the rates of reward leads to abrupt and complete adjustments to the new rates of reward at a short latency following the change in programmed rates.

No Effect of Immediately Preceding Interreward Intervals

If the rats based their estimates of the relative rates of reward simply on the ratio of the two most recent interreward intervals, one from each side, then their expected stay durations would change almost immediately, as we have observed. However, their expected stay durations would also covary with the large random fluctuations in the ratio seen during periods when the relative rate of reward is, in fact, constant. Their behavior would track the noise in the input. To see whether it did, we compared the distributions of stay durations observed following different ratios of the two most recent interreward intervals. This comparison was made under steady state conditions, that is, during all those portions of the relative condition sessions where the programmed rates of reward were fixed at some particular ratio.

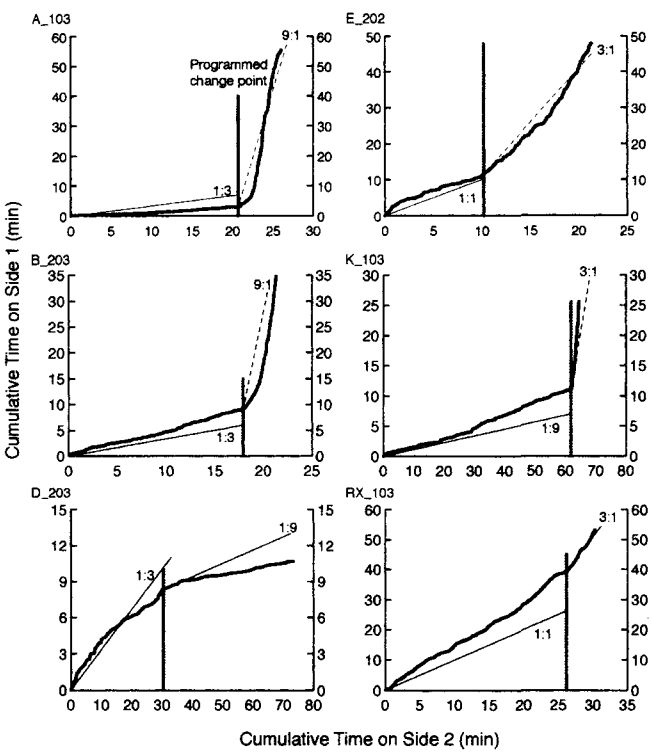

Figure 11. Cum-cum functions from the third or fourth sessions of the relative condition. Subject and session number are indicated at the upper left corner of each panel. The total number of unsignalled midsession changes in relative rates of reward that a subject had experienced (including the Phase 1–Phase 2 transition) is equal to the last two digits of the session number plus two. The subjects with session numbers in the 200 range had extensive experience with changes in the overall rates of reward; the subjects with session numbers in the 100 range had no such experience, at this point in the protocol.

To obtain populations of stay durations that were preceded by different ratios of interreward intervals, we grouped stays on the basis of binned interreward intervals. For each subject, we pooled the data from all those portions of a session with a given programmed ratio of reward rates. Within each of these data sets, we grouped stay durations on the basis of the immediately preceding interreward intervals on the two sides, with the interreward intervals themselves binned into intervals 0.2 log units wide. Binning the interreward intervals was necessary because few intervals are exactly the same when measured to the millisecond, and so, an exactly specified ratio rarely recurs. We excluded rewarded visits from the data because it was unclear how to define the immediately preceding interreward interval on a given side when there was a reward in the middle of a stay. Because the grouping of stays was determined by pairs of preceding interreward intervals—one from each side—the resulting groups formed cells in a table. The columns of the table were defined by the duration of the preceding interreward interval on one side, whereas the rows were defined by the duration of the preceding interreward interval on the other side.

If the ratio of the immediately preceding interreward intervals affected the expected duration of a stay—or any property of the distribution of stays—then the distributions of stays from cells defined by different immediately preceding ratios of interreward intervals would be different. To compare the distributions from cells that differed in this way, we made log survivor plots of the stays in the different cells. These plots were invariably almost superimposable, no matter how different the ratio of preceding intervals defining two cells (Figure 12). Stays that followed an unusually long interreward interval on Side 1 and an unusually short interreward interval on Side 2 had the same expected duration (and, indeed, the same distribution) as stays that followed the inverse ratio of interreward intervals. Thus, our subjects did not base their behavior on maximally local and maximally recent estimates of the expected intervals between rewards.

This result contradicts the conclusion of Mark and Gallistel (1994), who argued that stay durations did covary with the ratio of the immediately preceding interreward intervals. They did not, however, actually measure stay durations and interreward intervals because their system could not record event times. Their system logged total times on each side and numbers of rewards on each side within consecutive narrow windows defined by their software. These windows often included only one or two stays on each side. Because there are complex mutual dependencies between the different totals within such windows and significant autocorrelations (sequence dependencies) from window to window, we suspected that the strong correlations between the window-by-window ratios of the time totals and the window-by-window ratios of the reward totals might have been an artifact of the windowing imposed by their data recording system. We confirmed this by artificially windowing the data from the present experiment. We had the computer apply successive windows to the data series, totaling time and rewards within each window so as to simulate the manner in which the older apparatus converted the stream of events into window-by-window time and number totals.

The ratios of the totals from our artificial windows exhibited the correlations reported by Mark and Gallistel (1994), although the analysis reported above of the same data indicated no effect of the

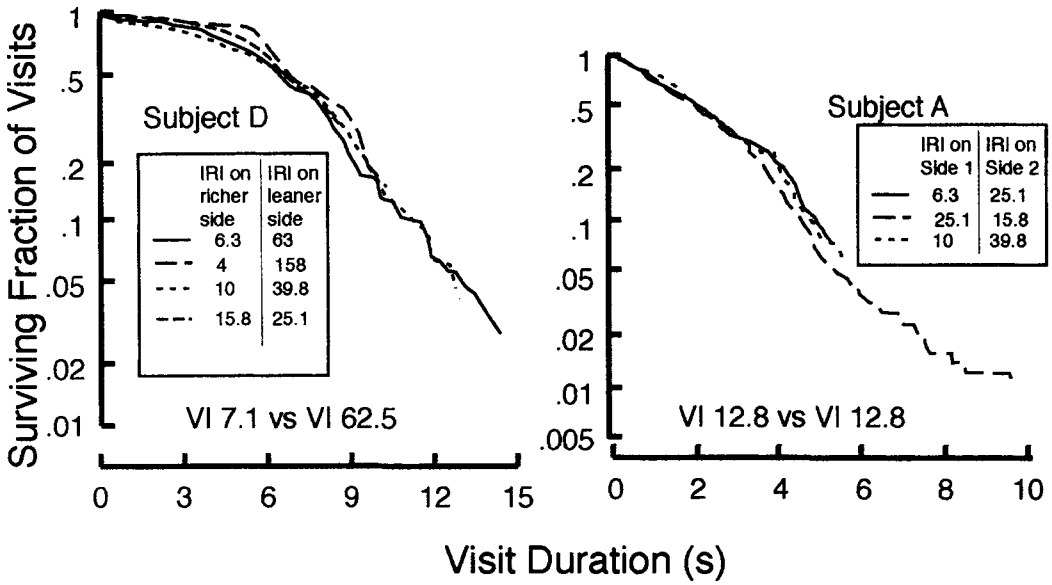

Figure 12. Log survivor plots for subpopulations of stays picked out on the basis of the durations of the immediately preceding interreward intervals on the two sides. The key gives the longer limits for each of the two interreward-interval bins used to pick out each subpopulation, one bin for each side. The shorter limit for each bin was 0.2 log units less than the longer limit. Stays that included a reward were excluded. Note that the distribution of stay durations is not affected by the durations of the preceding two interreward intervals nor by their ratio.

immediately preceding interreward intervals on the expected durations of stays. Thus, the findings that led Mark and Gallistel (1994) to argue for a dependence of stay duration on a maximally local sample of the interreward intervals appears to have been an artifact generated by the totaling of correlated times and response numbers within very narrow temporal windows.

Spontaneous Changes in Expected Stay Duration

Sometimes a change in behavior anticipates evidence of a change in the relative rates of reward. This could imply rodent clairvoyance, but this is not a hypothesis we seriously entertain. Another explanation for the relatively rare occasions on which the rat's adjustment precedes the appearance of objective evidence for a change in rates of reward is that there are spontaneous changes in the rats' expected stay durations, changes that occur in the absence of changes in the programmed rates of reward.

There is an example of a spontaneous purely transient change in the first-transition behavior of Subject A. If one looks closely at the cum-cum function for this subject on that session (upper left panel of Figure 2), one sees that almost immediately after the programmed change in the relative rates of reward, there happened to be an unusually prolonged stay on Side 1. This appears as a short vertical jump upward in the cum-cum function. This long stay on Side 1 is unlikely to have been a reaction to the change in the experienced rates of reward because it is in the wrong direction. This long stay made the cum-cum function temporarily steeper, whereas the asymptotic adjustment provoked by the programmed change was a decrease in the slope of the cum-cum function. As a consequence of this one stay, when the strength of the evidence for a change in behavior is plotted together with the strength of the evidence for a change in rate of reward (as in Figure 3), evidence for a change in behavior appears to develop well before the evidence for a change in rate (the initial upward excursion of the thick line in upper left panel of Figure 3), giving the appearance of clairvoyance.

Spontaneous and relatively enduring changes in the ratio of the rat's expected stay durations also occur (Figure 13). These are manifest in clear inflection points that occur well away from the programmed changes in the relative rates of reward. These spon-

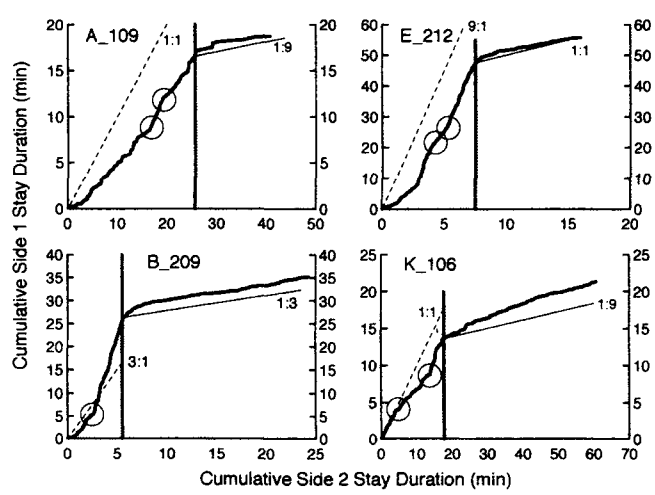

Figure 13. Examples of spontaneous abrupt adjustments in the subject's time-allocation ratio (circled). These were common once subjects had experienced many changes in relative rate.

taneous changes in the slope of the cum-cum function are sometimes in the direction of better matching, but they are not infrequently in the opposite direction: The slope of the cum-cum function is closer to the slope of the matching line before the change than after the change. In the Discussion we suggest a model for the decision process underlying matching that gives a causal explanation for these seemingly spontaneous enduring changes in the ratio of the expected stay durations.

Overadjustments

Not infrequently the abrupt adjustments in the relative condition were overadjustments; the change in the ratio of expected stay durations was greater than the change in the ratio of reward rates. Examples are A_117 in Figure 6 and A_103, D_203, and K_103 in Figure 11. This does not appear to be the kind of overshoot that one sees in an underdamped feedback process, which overshoots the asymptote at first and converges on it only after diminishing oscillations about it. These overadjustments are often made almost immediately after the programmed change in the relative rates of reward and then persist throughout the remainder of the semisession.

Discussion

Implications for the Law of Effect

The behavior observed in instrumental or operant conditioning experiments is widely assumed to develop through a feedback process in which behaviors are selected on the basis of their (subjective) consequences (see Williams, 1988, for review). As Schmajuk (1997, p. 149) put it, "During operant conditioning, animals learn by trial and error from feedback that evaluates their behavior but does not indicate the correct response." Rigorous formulations of this idea are given in the reinforcement learning literature in robotics and artificial intelligence (Mahadevan & Connell, 1992; Sutton & Barto, 1998) and in related neurobiologically oriented literature on learning (e.g., Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). A common way to implement the idea is to have reinforcement history predict the expected value of each contemplated action. Another way, less rigorously formulated so far, is to have the reinforcement history determine the strengths of stimulus–response associations (Lea & Dow, 1984). In either case, behavior is adjusted by the feedback effects of prior behavior until a reward function is maximized or returns are equated, a process we have described as hill climbing.

The principle involved—more frequently choosing in the future whatever course of action has produced a higher return in the past—would seem to be central to the rational decision making generally assumed in economic theory. However, our results are not consistent with models that implement this principle. If this kind of model is understood to be what is meant when one says that behavior has been instrumentally conditioned, then Heyman (1982) was right: Matching is not conditioned behavior; it is unconditioned behavior; that is, it is elicited by the animal's experience of the income ratio, independently of the returns it has experienced. The adjustments we report occur too soon and go to completion too rapidly to be produced by a hill-climbing process driven by changes in returns.

Heyman (1982) showed that the relation between the ratio of a subject's expected stay durations and the overall return it experiences is a weak one. The adjustment from ratios well away from matching to matching ratios increases the subject's overall return by only a few percentage points. This means that the hill being climbed has a very shallow gradient if the process is driven by changes in the overall rate of return. If, instead, the system is searching for the stay duration ratio that equates returns from the two sides (Herrnstein & Prelec, 1991), then the hill is steeper because the return from a given side is almost inversely proportional to the time invested in that side. However, the return from a side varies enormously from visit to visit—see middle panels in Figure 14. Because returns are so noisy, the effect on returns of a change in rates of reward takes some while to become evident. This means that a subject must wait some while before it can determine whether it has made its returns more or less equal by changing its expected stay durations. Moreover, it must adjust and then wait to see the consequences several times to find the new return-equating locus in behavioral space. The central idea behind the law of effect in operant conditioning is that reinforcement evaluates the behavior just performed but does not tell the animal what to do next (see the Schmajuk, 1997, quote above). We find, however, that the rat has often completed its adjustment to the new rates of reward before the change in rates of reward has had a measurable effect on its returns.

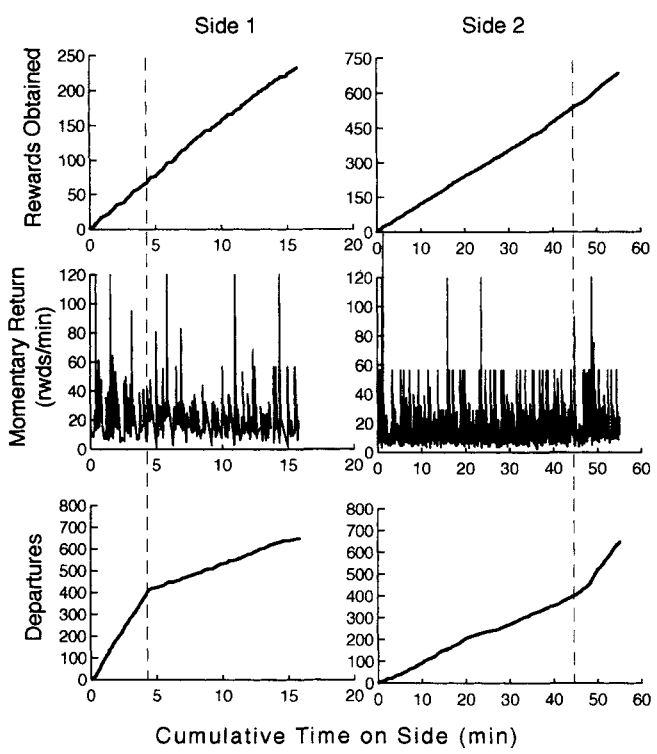

Figure 14. Top panels: Side-specific cumulative records of reward. Middle panels: Momentary returns (= 1/time on side since last reward on side) as a function of time on side. Bottom panels: Cumulative number of departures as a function of cumulative time on side. The slopes of these curves are the leaving rates. The dashed vertical lines indicate the point in the session at which the programmed change in the rates of reward occurred. Data from Subject A, Session 108.

The top panels in Figure 14 show the cumulative returns for the two sides, the numbers of rewards obtained as functions of the cumulative times spent on each of the sides. The slopes of these two lines are the average returns. They are approximately the same because matching equates returns. This fact is the basis for the melioration theory (Herrnstein & Prelec, 1991), according to which the subject allots ever more time to the side with the higher return until the returns from the two sides become equal.

The middle two panels of Figure 14 show the momentary returns on each side—the inverses of the successive interreward intervals, where those intervals are measured in time on side, not session time. These momentary returns may be thought of as the discrete derivatives of the cumulative records—discrete because they are defined only at the moments when rewards are delivered. As may be seen in Figure 14B, reward-by-reward returns are extremely variable, which makes it difficult to determine from only a few such returns whether there has been a change in the average return. The bottom two panels of Figure 14 show the number of departures as a function of the time on each side. The slopes of these functions are the leaving rates.

The dashed lines extending vertically across all three pairs of panels in Figure 14 indicate the point at which the programmed rates of reward changed. Notice the abrupt changes in the leaving rates that occur immediately after this change in reward rates (the changes in slopes seen in Figure 14C). The changes in leaving rates are unlikely to have been produced by changes in the experienced returns because there are no discernible changes in the returns (top and middle pairs of panels). Had the subject not adjusted its leaving rate, changes in the expected returns would eventually have become manifest. However, the subject adjusted to match the new relative rates before their effect on its expected returns emerged from the noise.

We confirmed the lack of significant perturbations in the experienced rates of return using the frequentist algorithm described later (see Figure 17) to compute after each reward the log of the odds that the expected return had changed up to that point. We compared this with the odds that the rat's expected stay durations had changed. Within 12 min following the programmed change in the rates of reward, the odds that the leaving rates had changed exceeded 1,000,000:1 on both sides (top panel in Figure 15), whereas the odds that there had been changes in the rates of return were on the order of 10:1 or less—corresponding to logits with absolute values less than 1 (bottom panel in Figure 15). These odds were no greater than those that occurred several times through random fluctuations earlier in the session, without provoking any noticeable behavioral adjustments. Thus, the change in leaving rates is unlikely to have been caused by a change in experienced returns because it occurred when there was no statistically meaningful evidence of a change in returns.

In a hill-climbing account like melioration (Herrnstein & Prelec, 1991), the subject does not know the ratio of expected stay durations that will equate its returns from the two sides until it has found it by (possibly guided) trial and error. It must try first one stay-duration ratio, determine the effect of that ratio on its returns, then adjust the ratio, and again determine the effect on its returns, and so on, until it finds the ratio that equates the returns. In this kind of model, the adjustment to a new ratio of rates of reward must be mediated by a sequence of relatively stable (or slowly changing) stay duration ratios, with successive ratios producing

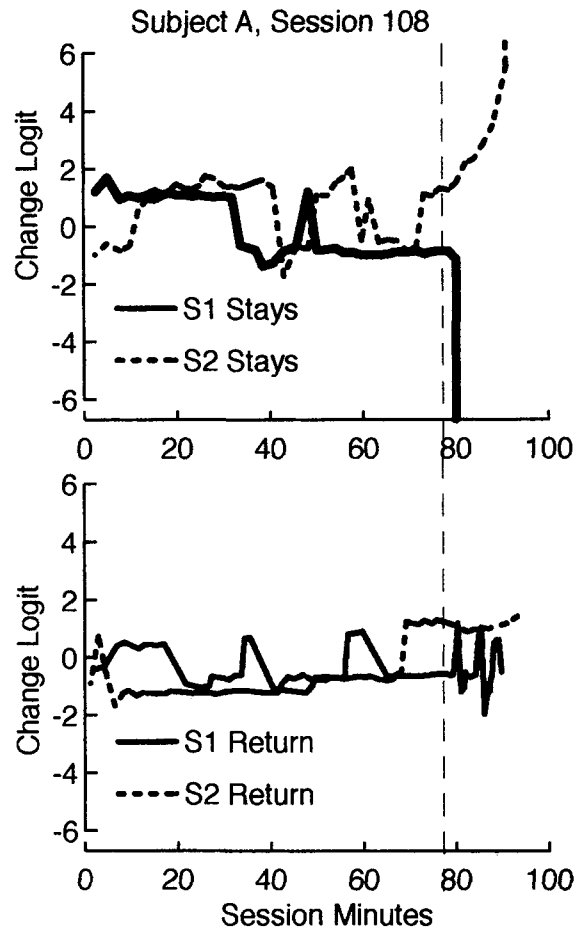

Figure 15. The log of the odds of a change in returns (bottom panel) and stay durations (top panel) as a function of session minutes for the session used to make the plots in Figure 14. A logit with absolute value of 6 indicates odds of 1,000,000:1 against the null hypothesis that there has been no change in rate. The vertical dashed line indicates the time at which the programmed rates of reward changed.

rates of return measurably different from the preceding rates. In fact, however, the adjustment may be completed before there is a measurable effect of the new rates of reward on the rates of return, which is why there is no perturbation in the slopes of the cumulative records at the top of Figure 14.

As Figure 14 shows, a change in the rates of reward may be evident in the experienced interreward intervals before it is evident in the experienced returns because random variation in visit durations may temporarily mask the effect on returns of a change in rates of reward. Suppose, for example, that a decrease in the rate of reward on Side 1 and an increase on Side 2 happen to coincide with a string of shorter visits on Side 1 and longer visits on Side 2 (shorter and longer relative to the respective expectations). During the sequence of fortuitously short visits to Side 1, the intervals between experienced rewards (measured in session time) are longer than usual because of the change in the rate of reward. The effects of these longer interreward intervals on the returns experienced is, however, masked by the fortuitous shortness of the visits because the returns experienced are the reciprocals of the

cumulative times spent on that side between rewards, and these times happen to be shorter than usual. The effect of the increased rate of reward on Side 2 is similarly masked; the subject experiences unexpectedly short interreward intervals but no increase in returns because of its fortuitously lengthier visits.

The abrupt and relatively enduring overadjustments often seen in our cum-cum records are also inconsistent with a hill-climbing model. A guided hill-climbing model, like the Gauss-Newton method used in many nonlinear curve-fitting algorithms, might show an overshoot. Guided search algorithms determine what to try next on the basis of the magnitude and direction of the change produced by the previous adjustment. A guided hill-climbing process might produce a transient overshoot, but the adjustment back toward the optimal location should occur as rapidly or more rapidly than the adjustment that culminated in the overshoot.

The spontaneous, relatively enduring adjustments away from matching are also inconsistent with a hill-climbing model. Such a model never leaves the hill top for any length of time, once it has come to rest there. In short, we believe that a process that selected stay duration ratios on the basis of the effects that different ratios have on experienced returns could not produce many features of our data.

Learning to Learn?

The increasingly rapid adjustments to frequently experienced changes that we observed are reminiscent of learning to learn phenomena, particularly the rapid reversals in discrimination learning that one sees if subjects are given repeated reversal training (Buytendijk, 1930; Dufort, Gutman, & Kimble, 1956; Krechevsky, 1932; North, 1950). In the just-cited experiments, subjects that repeatedly experienced reversal of the reinforcement contingency rapidly learned to reverse their choice when they encountered yet another reversal of the reinforcement contingency. However, the crucial variable, in the present case at least, was not whether subjects had learned to make rapid adjustments but rather the duration of the period of stability immediately preceding a change. After our subjects had (seemingly) learned to make rapid adjustments to changes in the relative rates of reward, they once again took a long time to adjust to a change that came after a long period of renewed stability. It appears not to be a matter of learning to adjust rapidly. Rather, it appears that the strategy for adjusting to a perceived change in the rates of reward takes into account the duration of the preceding period of stability. When the preceding period of stability is short, the strategy dictates immediate, full adjustment. When it is long, the strategy dictates a slower, more cautious adjustment. It should be noted in this connection that the latency to the onset of the adjustment process is about as short in the case of the slow adjustments as it is in the case of the rapid adjustments. It is the time course of the adjustment that changes, not the time that it takes to initiate it.

The conclusion that subjects adjust very rapidly when changes are frequent must be qualified by the observation that when the changes are so frequent that the animal encounters each condition only very briefly, it may cease to adjust to the changing conditions and average across conditions (e.g., Dreyfus, 1991, some conditions).

Spontaneous Recovery

Also notable are the reversions to the status quo ante that occur at the beginning of the next few sessions when there has been a midsession change in the relative rates of reward following a long period of stability (Figure 5; see also Mazur, 1996). This phenomenon is strongly reminiscent of the spontaneous recovery of a conditioned response, which is seen at the beginning of a new session, following a session in which the response was extinguished by repeatedly withholding reinforcement (for review, see Bouton, 1993b). If this reversion to the status quo ante is taken as one and the same phenomenon as spontaneous recovery, then it places an additional constraint on models of that phenomenon. The subjects must in effect remember the previous strengths of their conditioned responses, the strengths that they had before the rates of reinforcement were changed. One can no longer think simply in terms of an excitatory association, created by the original excitatory conditioning, and a competing inhibitory association, created by the extinction experience. A model in which the current strength or value of a response is determined by running averages of the excitatory (reinforcement) and inhibitory (nonreinforcement) effects of the responses made does not seem able to capture what is going on.

Elements of a Feed-Forward Model

In a feed-forward model, the mapping from the subject's representation of the conditioning situation to its behavior is immutable. There is no selection by consequences, hence there is no need to wait until changes in behavior have observable consequences. Changes in the subject's behavior are the result of changes in the values of the internal variables that constitute its representation or perception of its situation. In conditioning experiments, the relevant variables are primarily the temporal parameters of the experimental protocol (Gallistel & Gibbon, 2000).

We assume that to model matching behavior, we need primarily to specify two perceptual–representational processes and to specify the response strategies that translate the percepts into behavior. The first perceptual process estimates the current rates of reinforcement. The second detects a change in rates of reinforcement. The perception of a change causes the first process to reestimate the rates. The response strategies describe how the two percepts—perceived rates and perceived changes in rates—are translated into behavior.

Perceiving and responding to stable rates. Following Gallistel and Gibbon (2000), who built on earlier work by Myerson and Miezin (1980) and others (Heyman, 1982; Pliskoff, 1971; Staddon, 1977), we assume that the perception of the current rate of reinforcement at each location (on each lever) is the reciprocal of the arithmetic mean of a small sample of interreward intervals. We assume that perceived rate combines multiplicatively with subjective reward magnitude to determine subjective income (Leon & Gallistel, 1998). Under stable conditions (no recently perceived change in the rates of reward), perceived incomes translate into observed stay durations through a stochastic stay-terminating decision process. The stay-terminating process is assumed to be intrinsically stochastic rather than deterministic because after a minimum interval, the momentary likelihood of stay termination is independent of the duration of the stay (Figure 1, see also Heyman, 1982, Figures 3 and 4, and Gibbon, 1995, Figure 4). The expected

duration, $E(d_i)$, of a stay on Side i, is the reciprocal of the leaving rate on that side; the greater the leaving rate, the shorter the expected duration of a stay.

In the light of our present findings, we assume that the expected stay durations on the two sides are determined by the following constraints:

$$E(d_1)/E(d_2) = \hat{H}_1/\hat{H}_2 \tag{1}$$

and

$$\frac{1}{E(d_1)} + \frac{1}{E(d_2)} = a(\hat{H}_1 + \hat{H}_2) + b, \tag{2}$$

where $\hat{H}_i$ is the subject's estimate of the income on Side i. Equation (1) says that the ratio of the expected stay durations for two sides must equal the ratio of the experienced incomes. Equation (2) says that the sum of the leaving rates on the two sides is a linear function of the sum of the incomes. Thus, the greater is the combined income, the higher are both leaving rates and the shorter are the expected stay durations.

In an earlier formulation (Mark & Gallistel, 1994), the sum of the leaving rates was assumed to be simply proportional to the sum of the incomes. However, data from the phase of the current experiment in which we varied overall rates of reward show that the relation is in fact nonlinear (Figure 16). Leaving rate saturates as the overall reward rate becomes high. This will necessarily be the case when there is a minimum stay duration, as there appears to be (Figure 1). The subject does not, so to speak, start to consider leaving until it has been on a side some minimum amount of time. After that, its momentary likelihood of leaving is more or less constant. Because leaving rates appear to saturate, we calculated regression lines for all three levels of overall reward rate and for the two lower levels only—to see whether the line for the two lower levels would have an intercept significantly different from zero. In most cases it did. Thus, even if we assume that at lower overall rates of reward, the relation between overall reward rate and leaving rate is approximately linear, it is not one of simple proportionality. It will, however, approximate a proportional law at low enough rates of reward, where the influence of the nonzero intercept becomes negligible.

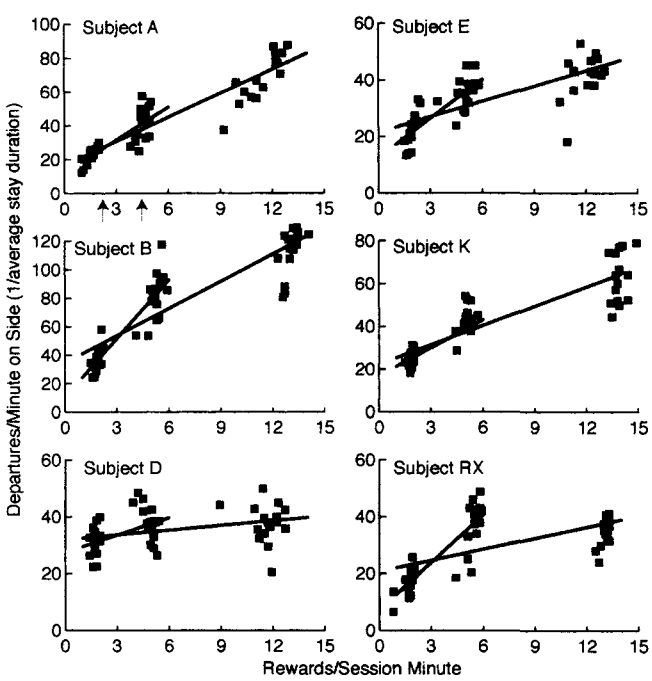

Figure 16. Overall leaving rates as a function of the overall reward density in the phase where the scheduled reward rates on the two levers were equal and the overall reward rate changed at the beginning of each session and again somewhere in the middle 80 min of the session (Phase 3 or 4, depending on the subject). The abscissa is the actually experienced reward density, not the programmed density. These densities (overall reward rates) cluster at three levels because there were only three levels of programmed overall rate. One regression line has been fitted to the data from all three levels of overall reward rate, whereas another has been fitted only to the data from the two lower levels.

The constraints given by Equations (1) and (2) uniquely determine leaving rates, given the subject's estimates of the incomes. The expression for the leaving rates is obtained by solving Equations (1) and (2) simultaneously to obtain the following:

$$\frac{1}{E(d_1)} = a\hat{H}_2 + b\frac{\hat{H}_2}{\hat{H}_1 + \hat{H}_2} \quad \text{and} \quad \frac{1}{E(d_2)} = a\hat{H}_1 + b\frac{\hat{H}_1}{\hat{H}_1 + \hat{H}_2}. \tag{3}$$

Equation (3) is an example of a mapping from a subject's perception of its situation (the expected incomes) to its pattern of behavior (average stay durations).

Responding to a perceived change. We suggest that when rates of reward change frequently, the subject's response to a perceived change in rates of reward is to make a new small sample of the interreward intervals immediately following the point at which it perceives the change in reward rate to have occurred. When changes are expected because they have recently been frequent, it immediately uses the rate estimates based on these new samples as the sole determinants of its stay durations.

On this model, spontaneous but enduring small changes in the ratio of the expected stay durations occur whenever the subject erroneously perceives a change in a rate of reward. The erroneous perception of a change in rate causes it to reestimate the rates. Small sample variability tends to make the new estimates differ appreciably but unsystematically from the old estimates. This would produce spontaneous but relatively enduring changes in the ratio of the expected stay durations, which are a clear feature of our data (Figure 13).

When there has not been an unequivocal change in rates of reward for a long time, then the subject does not immediately use the new samples as the sole determinants of its stay durations. Rather, the determinants of its new stay durations represent a compromise between the previous rates of long duration and the newly perceived rates, which have only been in effect for a short while. We do not attempt to specify the quantitative form of this compromise between the results of extended past experience and short but recent experience.

The reversions to the status quo ante that we and others observe imply that estimated incomes prior to perceived change points are not forgotten. They are kept in memory and may regain control of behavior under some circumstances. One such circumstance is at the beginning of a new session following a session in which there was an unprecedented or highly unusual change in the rates of reinforcement. In such a situation, it is inherently ambiguous

which is the better predictor of the incomes to be encountered at the start of a new session—those that prevailed during the latter part of the preceding session or those that prevailed for many sessions up to and including the beginning of the immediately preceding session.

A complete feed-forward model will have to have decision rules to deal with the case in which past sessions give ambiguous indications of what to expect in the new session. Devenport (1998) and collaborators (Devenport, Hill, Wilson, & Ogden, 1997) have proposed a temporal weighting rule, whereby animals weight conflicting experiences on the basis of their relative extent and recency. The extensive literature on spontaneous recovery after changes in temporal context may be taken as further evidence that the relative duration and recency of the experiences on which income estimates are based are themselves important determinants of behavior (Bouton, 1991, 1993a, 1993b).

Perceiving a change. In modeling the change-detecting process, we sought a simple real-time calculation yielding a decision variable sensitive enough to the information in the sequence of rewards to make the subject approximate an ideal detector. Detecting a change in a random rate is equivalent to detecting a change in the slope of the cumulative record, which is the plot of the number of events as a function of elapsed time (Figure 17). Assume that the observation of events begins at time t_0, which is the time of occurrence of the event number, n_0. Let $t > t_0$ be a subsequent point in time, $T = t - t_0$ be the interval of observation, and $N > 0$ be the number of events observed in that interval, including the event, if any, at t, but not including the event at t_0. If the rate has been constant, then the cumulative record approximates a straight line, with slope N/T (see, for empirical examples, the top panels in Figure 14).

If the rate has changed somewhere within the interval of observation, then the cumulative record will have an inflection point where it changed. Let t_c be the time at the inflection point and n_c the event count. If there is an inflection point in the record, then its location may be estimated to be the point at which the record deviates maximally from a straight line (d_{max} in Figure 17). Because the cumulative record is incremented in discrete steps, this point always coincides with an event. Let $T_a = t - t_c$ be the interval since the putative inflection point and N_a be the number of events observed after that point.

On the null hypothesis that the events are randomly distributed in time, the probability, p_e, that any one of the events (ignoring event order) falls in the interval T_a is T_a/T. The probability P_f of observing N_a or fewer events is given by the cumulative binomial probability function, as is the probability P_m of observing N_a or more events. When the number of events in T_a is approximately the expected number, then the ratio P_f/P_m is approximately unity and the log of this ratio is approximately 0. As the observed number of events since the putative time of change becomes improbably low, the ratio becomes very small, and its log approaches minus infinity. As the observed number becomes improbably high, the ratio becomes very large, and its log approaches infinity. The absolute value of the log of this ratio (the logit[2]) is our proposed subjective measure of the strength of the evidence that there has been a change in rate. When this quantity exceeds a critical value, the subject perceives a change. When the subject perceives a change in a rate, it truncates the data at the moment the change is perceived to occur (the moment t_c in Figure 17). The data on which the next perception of a change in rate is based are only those after this moment.

To check whether our model for the computation of the decision variable underlying change perception approximates the behavior of an ideal detector, we plotted our measure against the real-time odds for a change calculated by the Bayesian formula in the Appendix (A13a/A13b). Figure 18 compares the values of our measure with the value from the Bayesian calculation when applied to representative reward and stay duration data series. To a good approximation, the Bayesian measure of the strength of the evidence for a change and the frequentist measure provided by the above described algorithm differ only by a scaling factor. Thus, our model of the change-detecting process satisfies the requirement that it approximate the performance of an ideal detector. We used this algorithm to compute the evidence for a change in expected stay durations and the evidence for a change in expected returns in Figure 15.

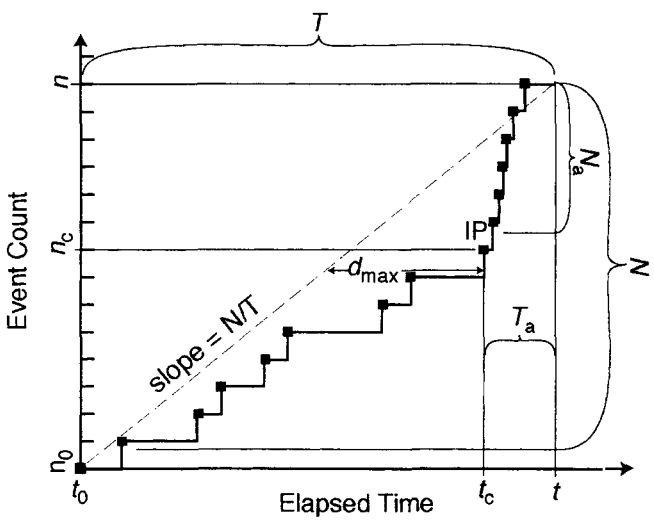

Figure 17. A cumulative record of events (cumulative number of events vs. time on the clock). So long as the average rate of event occurrence remains constant, this function will approximate the dashed straight line. When the rate changes, there is an inflection point (IP), with coordinates $\langle t_c, n_c \rangle$.

Conclusions

The short latency at which subjects begin to adjust their expected stay durations in reaction to changes in the relative rates of reward and the abruptness of these adjustments when such changes are frequent imply that matching behavior is not the result of a learning process that selects behaviors on the basis of their consequences. The locus of reinforcement's effect is not in the mapping from perceived situations to actions nor in the mapping from actions to the amounts of reward they are expected to produce;

[2] The logit is usually defined to be the ratio of two complementary probabilities. Our ratio is between two overlapping probabilities, which therefore do not sum to 1. We use overlapping probabilities because the resulting measure is better behaved when the expected and observed numbers of events are the same and near or equal to zero. Away from unity or when the expected number of events is $\gg 0$, our ratio approximates the usual ratio.

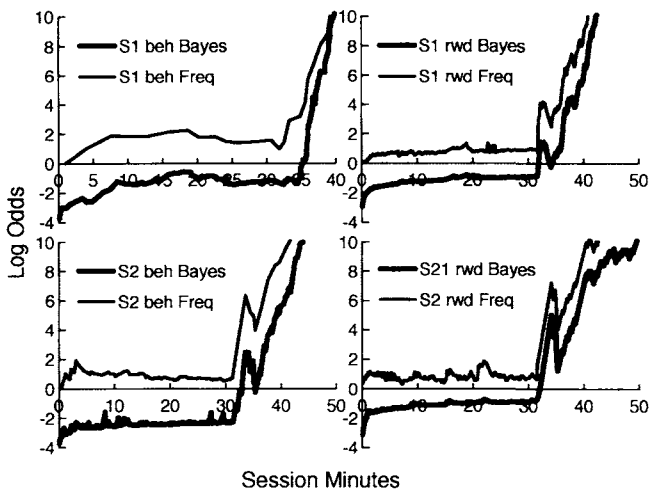

Figure 18. The results from our frequentist model of the change-detecting process (Freq curves) compared with the results from the Bayesian formula (Bayes curves, calculated with Formula A13 in the Appendix), using data on the changes in the experienced rates of reward and answering changes in the rat's stay durations from Session 110 of Subject A. beh = behavior; rwd = reward.

rather, it is in the subject's representation of income histories. That representation determines its behavior via a seemingly immutable decision process (Herrnstein, 1991). Feedback effects may be relevant only insofar as they alter this representation.[3]

A pigeon or rat learns to peck a key or press a lever for food reinforcement simply by observing the contingency between the illumination of the key or the appearance of the lever and the delivery of food, regardless of the effect its own behavior does or does not have on food delivery (Brown & Jenkins, 1968). It responds to the perceived contingency even when its response causes the omission of the reward it would otherwise receive (Williams & Williams, 1969). Thus, the mere perception of a stimulus–reward contingency can elicit seemingly instrumental behavior directed toward the stimulus. We have now shown that the perception of the incomes to be expected elicits matching behavior. Thus, both the appearance of conditioned responses and their relative strengths may depend simply on perceived patterns of reward without regard to the behavior that produced those rewards.

These findings raise the question, under what circumstances is instrumental behavior formed through a process of selecting behaviors on the basis of their consequences? That the effects of an animal's past behavior are important determinants of its future behavior is beyond dispute. However, this influence of past effects on future behavior need not arise from the fact that the relevant experiences were consequences of the subject's behavior. The behavior may have served only to reveal aspects of the environment that the animal would not otherwise have experienced. Unless it samples a location, an animal cannot know what to expect there. However, the behaviorally important expectation need make no reference to the sampling behavior itself. The distinction between income and return is the distinction between an expectation whose computation does not depend on knowing what behavior produced the outcomes and an expectation whose computation does require such knowledge. We have shown that matching is driven by the former. This explains why matching is observed under group conditions where all subjects have observed the rates at which food is delivered to two different patches, but most subjects have not yet succeeded in obtaining food from one or both patches (Harper, 1982).

At least the following variables appear to be crucial determinants of matching behavior: (a) the currently prevailing incomes (amounts of reward per unit of session time), (b) whether there has been a recent change in the incomes, (c) the recency of that change, and (d) the duration of the period of stable incomes preceding the change.

The process that detects changes in rates of reward approximates an ideal detector, which implies that the subjective likelihood that there has been a change approximates the objective likelihood. To approximate an ideal detector of changes in rate, a subject must remember the sequence of interevent intervals leading to the present moment.

[3] The pure feed-forward view does not preclude feedback effects because the animal's behavior may affect the interreward intervals that it experiences. If, for example, the animal samples a given option only at longer and longer intervals, then it will necessarily experience rewards from that option only at longer and longer intervals, so the income from that option will go down. In cases where feedback from behavior to experience is demonstrably important, the question concerns the locus of the effects of this feedback. Is the locus a change in the experienced income? Or is it a change in the mapping between experienced stimulus situations and response outputs? The law of effect asserts the latter: It asserts that the effects of behavior alter the mapping from experienced stimulus situations to response outputs. This is also the assumption in the reinforcement-learning approach to complex decision making in artificial intelligence and comparable analyses in economic decision theory. In all of these different domains, it is assumed that the locus of the agent's adjustment is in his response strategy, that is, his decision process, rather than in his representation of the situation-defining variables (his model of the relevant aspects of the world).

References

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *Journal of the Experimental Analysis of Behavior, 36,* 387–403.

Bouton, M. E. (1991). Context and retrieval in extinction and in other examples of interference in simple associative learning. In L. Dachowski & C. R. Flaherty (Eds.), *Current topics in animal learning* (pp. 25–53). Hillsdale, NJ: Erlbaum.

Bouton, M. E. (1993a). Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science, 3,* 49–53.

Bouton, M. E. (1993b). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114,* 80–99.

Brown, P. L., & Jenkins, H. M. (1968). Autoshaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior, 11,* 1–8.

Buytendijk, F. J. J. (1930). Über das Umlernen [On reversal learning]. *Archiv der neerländiscen Physiologie, 15,* 283–310.

Davison, M., & McCarthy, D. (1988). *The matching law: A research review.* Hillsdale, NJ: Erlbaum.

Devenport, L. D. (1998). Spontaneous recovery without interference: Why remembering is adaptive. *Animal Learning and Behavior, 26,* 172–181.

Devenport, L., Hill, T., Wilson, M., & Ogden, E. (1997). Tracking and averaging in variable environments: A transition rule. *Journal of Experimental Psychology: Animal Behavior Processes, 23,* 450–460.

DeWeese, M., & Zador, A. (1998). Asymmetric dynamics in optimal variance adaptation. *Neural Computation, 10,* 1179–1202.

Dreyfus, L. R. (1991). Local shifts in relative reinforcement rate and time allocation on concurrent schedules. *Journal of Experimental Psychology: Animal Behavior Processes, 17,* 486–502.

Dufort, R. H., Gutman, N., & Kimble, G. A. (1956). One-trial discrimination reversal in the white rat. *Journal of Comparative and Physiological Psychology, 47,* 248–249.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review, 107,* 289–344.

Gamzu, E., & Williams, D. R. (1971). Classical conditioning of a complex skeletal response. *Science, 171,* 923–925.

Gibbon, J. (1995). Dynamics of time matching: Arousal makes better seem worse. *Psychonomic Bulletin and Review, 2,* 208–215.

Harper, D. G. C. (1982). Competitive foraging in mallards: Ideal free ducks. *Animal Behavior, 30,* 575–584.

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. (1991). Experiments on stable sub-optimality in individual behavior. *American Economic Review, 81,* 360–364.

Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives, 5,* 137–156.

Herrnstein, R. J., & Vaughan, W. J. (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits to action: The allocation of individual behavior* (pp. 143–176). New York: Academic.

Heyman, G. M. (1979). A Markov model description of changeover probabilities on concurrent variable-interval schedules. *Journal of the Experimental Analysis of Behavior, 31,* 41–51.

Heyman, G. M. (1982). Is time allocation unconditioned behavior? In M. Commons, R. Herrnstein, & H. Rachlin (Eds.), *Quantitative analyses of behavior: Matching and maximizing accounts* (Vol. 2, pp. 459–490). Cambridge, MA: Ballinger Press.

Heyman, G. M., & Luce, R. D. (1979). Operant matching is not a logical consequence of maximizing reinforcement rate. *Animal Learning and Behavior, 7,* 133–140.

Krechevsky, I. (1932). Antagonistic visual discrimination habits in the white rat. *Journal of Comparative Psychology, 14,* 263–277.

Lea, S. E. G., & Dow, S. M. (1984). The integration of reinforcements over time. In J. Gibbon & L. Allan (Eds.), *Timing and time perception* (Vol. 423, pp. 269–277). New York: Annals of the New York Academy of Sciences.

Leon, M. I., & Gallistel, C. R. (1998). Self-stimulating rats combine subjective reward magnitude and subjective reward rate multiplicatively. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 265–277.

Mahadevan, S., & Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence, 55,* 311–365.

Mark, T. A., & Gallistel, C. R. (1994). The kinetics of matching. *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 79–95.

Mazur, J. E. (1995). Development of preference and spontaneous recovery in choice behavior with concurrent variable-interval schedules. *Animal Learning & Behavior, 23,* 93–103.

Mazur, J. E. (1996). Past experience, recency, and spontaneous recovery in choice behavior. *Animal Learning & Behavior, 24,* 1–10.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16*(5), 1936–1947.

Myerson, J., & Miezin, F. M. (1980). The kinetics of choice: An operant systems analysis. *Psychological Review, 87,* 160–174.

North, A. J. (1950). Improvement in successive discrimination reversals. *Journal of Comparative and Physiological Psychology, 43,* 442–460.

Pliskoff, S. S. (1971). Effects of symmetrical and asymmetrical changeover delays on concurrent performances. *Journal of the Experimental Analysis of Behavior, 16,* 249–256.

Schmajuk, N. A. (1997). *Animal learning and cognition.* New York: Cambridge University Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Staddon, J. E. R. (1977). On Herrnstein's equations and related forms. *Journal of the Experimental Analysis of Behavior, 28,* 163–170.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning.* Cambridge, MA: MIT Press.

Williams, B. A. (1988). Reinforcement, choice, and response strength. In R. C. Atkison, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Steven's handbook of experimental psychology, 2nd ed.* (Vol. 2, pp. 167–244). New York: Wiley.

Williams, D. R., & Williams, H. (1969). Automaintenance in the pigeon: Sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior, 12,* 511–520.

Appendix

Optimal Estimation of Rat Behavior by the Observer and of Reward Rates by the Rat

A major goal of this article is to determine how well a rat responds to a change in reward rates compared with how well it could respond in principle. There are two parts to this problem. The first is to determine what we mean by in principle; that is, what is the fastest any observer could detect a change in reward rates, given that those rewards appear randomly. The second is assessing whether the rat detected the change; our only way of doing that is by observing the rat's behavior—specifically, the stay durations on one side of an experimental apparatus—which itself is a random process.

These two problems share a common feature in that both involve detection of a change in a random process. There is, however, one major difference. To assess when a rat changes its behavior, we can examine data over an extended period; that is, we can make our decision on the basis of knowledge of all stay durations. The rat, however, has a more difficult problem: it must make its decision in real time, only on the basis of reward times that it has seen so far, not on all reward times. Moreover, it must decide not only when a reward rate changed, but what it changed to.

We address first the problem of finding the time at which a rat changed its behavior given all stay durations; after that, we tackle the problem of estimating changes in reward rates in real time.

Determining When a Rat Changes Its Stay Durations, Given All the Data

To make inferences about changes in a rat's stay durations, we need a model for the statistics of those durations. On the basis of the experimental data (see main text, *Distributions of Stay Durations*), to a good approxi-

mation stay durations are exponentially distributed. Consequently, the times at which a rat leaves a side follow a Poisson process. Our problem, then, is to determine when a Poisson process undergoes a change in rate, given that we have access to all the data. We formulate this problem as follows: At time $t = 0$, a Poisson process is initiated, in which events are generated at rate r_0. At time t_c, the rate changes to r_1, and at $t = T$ the events stop. Given the event times, denoted $\mathbf{t} = t_1, t_2, \ldots$, we want to know when the rate changed from r_0 to r_1. Because we cannot know the exact time at which the rates changed, the best we can hope for is the probability distribution of t_c. In other words, we want to construct the conditional probability distribution $P(t_c|\mathbf{t})$.

Our strategy is to use Bayes' theorem to compute $P(t_c, r_0, r_1|\mathbf{t})$, the probability density of observing rates r_0 and r_1 before and after time t_c, respectively, then integrate over r_0 and r_1 to find $P(t_c|\mathbf{t})$. Bayes' theorem is

$$P(r_0, r_1, t_c|\mathbf{t}) = \frac{P(\mathbf{t}|r_0, r_1, t_c)P(r_0, r_1, t_c)}{P(\mathbf{t})}. \tag{A1}$$

We are using the potentially ambiguous notation that a probability distribution is specified by its arguments, but in all cases what we mean should be clear from the context.

To determine $P(r_0, r_1, t_c|\mathbf{t})$, we need expressions for the two terms in the numerator on the right hand side of Equation A1. Let us begin with $P(\mathbf{t}|r_0, r_1, t_c)$, the probability distribution of the event times given that the rate is r_0 when $t \leq t_c$ and r_1 when $t > t_c$. Because events are generated by a Poisson process, the probability of observing a particular set of event times, $\mathbf{t}$, depends only on the number of events that occurred before time t_c and the number that occurred after t_c. We denote those two numbers l_0 and l_1, respectively. Then, discretizing time into bins of width Δt, small enough so that each bin is overwhelmingly likely to contain at most one reward (we eventually let $\Delta t \to 0$), we have

$$P(\mathbf{t}|r_0, r_1, t_c) = (r_0\Delta t)^{l_0}(1 - r_0\Delta t)^{t_c/\Delta t - l_0}(r_1\Delta t)^{l_1}(1 - r_1\Delta t)^{(T-t_c)/\Delta t - l_1}. \tag{A2}$$

The quantity $(r_0\Delta t)^{l_0}$ is the probability of observing events occurring in l_0 time bins of width Δt, whereas $(1 - r_0\Delta t)^{t_c/\Delta t - l_0}$ is the probability of no events occurring in the remaining $t_c/\Delta t - l_0$ bins. Identical logic applies to the third and fourth terms in this expression. For small enough Δt, Equation A2 becomes

$$P(\mathbf{t}|r_0, r_1, t_c) \propto r_0^{l_0}e^{-r_0t_c}r_1^{l_1}e^{-r_1(T-t_c)}, \tag{A3}$$

where we used $(1 - a\epsilon)^{b/\epsilon+c} \to e^{-ab}$ in the limit $\epsilon \to 0$. The constant of proportionality that relates the left and right hand sides of Equation A3 is determined later.

To complete our expression for $P(r_0, r_1, t_c|\mathbf{t})$, we need the priors—the probability of leaving rates and the time at which a change occurred, both in the absence of any data. Although the true priors are known from the experiment, for this calculation it is sufficient to use flat priors, $P(r_0, r_1, t_c) =$ constant. This is because the conditional distribution, $P(\mathbf{t}|r_0, r_1, t_c)$, quickly becomes sharply peaked as soon as there is even a small amount of data, rendering the priors more or less irrelevant. With flat priors, we can insert Equation A3 into Equation A1 to yield

$$P(r_0, r_1, t_c|\mathbf{t}) \propto r_0^{l_0}e^{-r_0t_c}r_1^{l_1}e^{-r_1(T-t_c)}.$$

Integrating over both r_0 and r_1 from zero to infinity, we arrive at the expression for our desired quantity, $P(t_c|\mathbf{t})$,

$$P(t_c|\mathbf{t}) = \frac{1}{\Omega}\frac{l_0!}{t_c^{l_0+1}}\frac{l_1!}{(T-t_c)^{l_1+1}}, \tag{A4}$$

where the constant of proportionality, Ω, satisfies

$$\Omega = \int dt_c \frac{l_0!}{t_c^{l_0+1}}\frac{l_1!}{(T-t_c)^{l_1+1}}. \tag{A5}$$

In both Equations A4 and A5, l_0 and l_1 depend on t_c as well as the event times, $\mathbf{t}$. The integral in the latter equation must be done numerically, once the data is given. In practice, to avoid infinities in the $t_c \to 0$ and $t_c \to T$ limits, we set the lower and upper bounds on the integral to correspond to the first and last event times.

Equation A4 is the probability distribution of a change in reward rate given a complete set of event times, $\mathbf{t}$. We now address the problem of finding the probability that a change in rates has occurred, and what the new rate is.

Determining, in Real Time, When the Reward Rates Change

The problem faced by the rat is to collect the reward times—the data—up to a certain point in time and then generate a probability distribution for the actual rate of rewards given that data. We assume that rewards are generated according to a Poisson process (although this is not quite true in the experiments), so we can state our problem as follows: At time $t = 0$, rewards started appearing, and we have been observing those rewards between $t = 0$ and $t = t_f$. Rewards appear first at rate r_0 and, at time t_c, change to rate r_1. The reward rate can change at most once; that is, there is only one value of t_c. Of course, t_c may be greater than t_f, in which case the change in reward rate will not have been made during our observation time. Our goal is to construct the probability distribution for the reward rates before and after the change in rate.

Notice that the previous method cannot be applied directly; we cannot simply replace T by t_f in Equation A4. This is because there is a finite probability that the reward rate has not changed at time t_f, a possibility that was not taken into account in deriving Equation A4. Instead, we use the method described by DeWeese and Zador (1998), which casts the problem in terms of a hidden Markov model. The idea is as follows. As above, we start by discretizing time into small bins of width Δt. Let r_i be the rate in interval i and $s_i = \{0, 1\}$ be an indicator of whether ($s_i = 1$) or not ($s_i = 0$) a reward was left. Let x_i indicate whether or not the reward rate has changed: $x_i = 1$ indicates that there has been no change, and $x_i = 0$ indicates a change has taken place. This indicator is needed because rewards change at most once. The method outlined in DeWeese and Zador (1998) is to compute $P_i(r_i, x_i|s_{j\leq i})$—the probability that the reward rate is r_i and the rate has (or has not) changed, given observations of all rewards up to interval i—as a functional of $P_{i-1}(r_{i-1}, x_{i-1}|s_{j\leq i-1})$. We can then iterate, starting at $i = 0$, to find P_i for any index i. The subscript on the probability distribution reminds us that P_i and P_j are different if $i \neq j$.

In the experiments, the rewards follow (approximately) a Poisson process, and the decision to change the rewards is independent of reward times. This leads to two observations:

1. Whether a reward is left in interval i depends only on the reward rate, r_i, in that interval.

2. The reward rate, r_i, and indicator x_i, in interval i depends only on the reward rate and the indicator in the previous interval; that is, they depend only on r_{i-1} and x_{i-1}.

These observations allow us to determine how $P_i(r_i, x_i|s_{j\leq i})$ depends on $P_{i-1}(r_{i-1}, x_{i-1}|s_{j\leq i-1})$. We start with the well-known relation between joint probability distributions and conditional distributions,

$$P_i(r_i, x_i, s_{j\leq i}) = P_i(s_i|r_i, x_i, s_{j\leq i-1})P_i(r_i, x_i, s_{j\leq i-1}). \tag{A6}$$

Because s_i depends only on r_i (Observation 1), the first term on the right-hand side reduces to $P_i(s_i|r_i)$. The second term can be recast using Observation 2: In the absence of knowledge of s_i, r_i, and x_i depend *only* on r_{i-1} and x_{i-1}. Thus, the second term becomes

(*Appendix continues*)

$$P_i(r_i, x_i, s_{j\le i-1}) = \sum_{x_{i-1}} \int dr_{i-1} P_i(r_i, x_i | r_{i-1}, x_{i-1}) P_{i-1}(r_{i-1}, x_{i-1}, s_{j\le i-1}). \quad \text{(A7)}$$

Combining Equations A6 and A7 and again turning joint distributions into conditional ones, we find that

$$P_i(r_i, x_i | s_{j\le i}) = \frac{1}{Z} P(s_i|r_i) \sum_{x_{i-1}} \int dr_{i-1} P(r_i, x_i | r_{i-1}, x_{i-1}) P_{i-1}(r_{i-1}, x_{i-1} | s_{j\le i-1}), \quad \text{(A8)}$$

where Z is the normalization, chosen so that $\sum_{x_i} \int dr_i P_i(r_i, x_i | s_{j\le i}) = 1$, and we have dropped the subscripts on both $P(s_i|r_i)$ and $P(r_i, x_i | r_{i-1}, x_{i-1})$ because neither depends on the time interval.

To implement Equation A8 for a particular problem, all we need to do is write down expressions for $P(s_i|r_i)$ and $P(r_i, x_i | r_{i-1}, x_{i-1})$. Because rewards are left according to a Poisson process, the first of these may be written

$$P(s_i|r_i) = (1 - r_i\Delta t)\delta_{s_i,0} + r_i\Delta t \delta_{s_i,1}, \quad \text{(A9)}$$

where $\delta_{i,j}$ is the Kronecker delta: $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. To determine $P(r_i, x_i | r_{i-1}, x_{i-1})$, we need to know how reward changes are distributed. In the experiments described in the main text, the probability of a reward change is uniform over a fixed interval. Were we to strictly follow this prescription, the probability of a reward change would steadily increase. It is unlikely, however, that the rat has such a sophisticated model. Instead, we assume that the probability of a change is fixed in any interval at $\nu\Delta t$ provided no change has occurred ($x_i = 1$) and is zero if a change has occurred ($x_i = 0$). Thus,

$$P(r_i, x_i | r_{i-1}, x_{i-1}) = (1 - x_{i-1}\nu\Delta t)\delta(r_i - r_{i-1})\delta_{x_i,x_{i-1}} + x_{i-1}\nu\Delta t g(r_i)\delta_{x_i,0}, \quad \text{(A10)}$$

where $g(r_i)$ is the probability distribution of a new rate given there was a change in rates and, in this expression, $\delta(r)$ is the Dirac δ function. The first term on the right-hand side of Equation A10 ensures that r_i and x_i only change once ($x_i = 0$ if a change has taken place); the second term allows a change in reward rate, with probability $\nu\Delta t$, given that no change has taken place so far.

To derive an update rule, we simply insert Equations A9 and A10 into A8 and perform the integral over r_{i-1} and the sum over x_{i-1}. We do this in steps, starting with the integral and sum. Using Equation A10, we have

$$\sum_{x'} \int dr' P(r, x | r', x') P_{i-1}(r', x') = (1 - x\nu\Delta t)P_{i-1}(r, x) + \nu\Delta t g(r)\delta_{x,0} \int dr' P_{i-1}(r', 1).$$

Then, using this expression and Equation A9 for $P(s|r)$, Equation A8 becomes

$$P_i(r, x) = \frac{1}{Z}[(1 - r\Delta t)\delta_{s,0} + r\Delta t\delta_{s,1}]\left[(1 - x\nu\Delta t)P_{i-1}(r, x) + \nu\Delta t g(r)\delta_{x,0}\int dr' P_{i-1}(r', 1)\right].$$

Note that we have dropped the subscript i on r, x, and s, as these are now dummy variables. To lowest order in Δt, this expression simplifies to

$$P_i(r, x) = \frac{1}{Z}\delta_{s,0}\left\{P_{i-1}(r, x) - \Delta t\left[(r + x\nu)P_{i-1}(r, x) - \nu g(r)\delta_{x,0}\int dr' P_{i-1}(r', 1)\right]\right\} + \frac{1}{Z}\delta_{s,1}\Delta t r P_{i-1}(r, x), \quad \text{(A11)}$$

from which we find that $Z = \delta_{s,0}[1 - \bar{r}\Delta t] + \delta_{s,1}\bar{r}\Delta t$, where

$$\bar{r} \equiv \sum_x \int dr r P_{i-1}(r, x).$$

Inserting this expression for Z into Equation A11 and again working to lowest order in Δt, we arrive at

$$P_i(r, x) = \delta_{s,0}\left\{P_{i-1}(r, x) - \Delta t\left[(r - \bar{r} + x\nu)P_{i-1}(r, x) - \nu g(r)\delta_{x,0}\int dr' P_{i-1}(r', 1)\right]\right\} + \delta_{s,1} r P_{i-1}(r, x)/\bar{r}.$$

In the limit $\Delta t \to 0$, this may be cast as a differential equation,

$$\frac{\partial P(r, x)}{\partial t} = -(r - \bar{r} + x\nu)P(r, x) + \nu g(r)\delta_{x,0}\int dr' P(r', 1) + \sum_k \delta(t - t_k)P(r, x)\frac{r - \bar{r}}{\bar{r}}, \quad \text{(A12)}$$

where t_k is the time of the kth reward.

Finally, it is convenient to break Equation A12 into two equations, one for $P_0(r) \equiv P(r, 0)$ and one for $P_1(r) \equiv P(r, 1)$, yielding

$$\frac{\partial P_0(r)}{\partial t} = -(r - \bar{r})P_0(r) + \nu g(r)\int dr' P_1(r') + \sum_k \delta(t - t_k)P_0(r)\frac{r - \bar{r}}{\bar{r}} \quad \text{(A13a)}$$

$$\frac{\partial P_1(r)}{\partial t} = -(r - \bar{r} + \nu)P_1(r) + \sum_k \delta(t - t_k)P_1(r)\frac{r - \bar{r}}{\bar{r}}. \quad \text{(A13b)}$$

Equations A13a and A13b contain the probability distributions for the reward rates both before (P_1) and after (P_0) there has been a change in the reward rate. Initial conditions are chosen so that $P_1(r) = g(r)$ and $P_0(r) = 0$, where $g(r) = 1/r_{max}$ when $r \le r_{max}$ and 0 when $r > r_{max}$. The maximum reward rate, r_{max}, is chosen to be larger than any reward rate likely to be seen by the rat. The odds that there has been a change in rate is simply given by $\int dr P_0(r)/\int dr P_1(r)$.

Received July 31, 2000
Revision received May 17, 2001
Accepted May 18, 2001 ■