LINEAR-NONLINEAR-POISSON MODELS OF PRIMATE CHOICE DYNAMICS

Greg S. Corrado, Leo P. Sugrue, H. Sebastian Seung, and William T. Newsome

HOWARD HUGHES MEDICAL INSTITUTE,
STANFORD UNIVERSITY SCHOOL OF MEDICINE,
AND MASSACHUSETTS INSTITUTE OF TECHNOLOGY

The equilibrium phenomenon of matching behavior traditionally has been studied in stationary environments. Here we attempt to uncover the local mechanism of choice that gives rise to matching by studying behavior in a highly dynamic foraging environment. In our experiments, 2 rhesus monkeys (*Macacca mulatta*) foraged for juice rewards by making eye movements to one of two colored icons presented on a computer monitor, each rewarded on dynamic variable-interval schedules. Using a generalization of Wiener kernel analysis, we recover a compact mechanistic description of the impact of past reward on future choice in the form of a Linear-Nonlinear-Poisson model. We validate this model through rigorous predictive and generative testing. Compared to our earlier work with this same data set, this model proves to be a better description of choice behavior and is more tightly correlated with putative neural value signals. Refinements over previous models include hyperbolic (as opposed to exponential) temporal discounting of past rewards, and differential (as opposed to fractional) comparisons of option value. Through numerical simulation we find that within this class of strategies, the model parameters employed by animals are very close to those that maximize reward harvesting efficiency.

Key words: matching, choice, decision theory, neuroeconomics, reward, LNP models, hyperbolic discounting, eye movements, monkey

In this journal, over a decade before the birth of the first two authors of this article, Richard Herrnstein published a simple observation about the choice behavior of animals in a key-pressing task: ‘‘The relative frequency of responding on a given key closely approximated the relative frequency of reinforcement on that key’’ (Herrnstein, 1961). If, for example, a pigeon received two thirds of its food-pellet rewards for pressing a particular key, the pigeon came to press that key two thirds of the time. By 1970, this observation had grown into a general law relating choice behavior to reward history, now commonly referred to as Herrnstein's Matching Law, which he also published here in the most widely cited scientific article in *JEAB*'s history (*JEAB*, 1993). The matching law asserts that:

$$\frac{r_k}{\sum r_i} = \frac{c_k}{\sum c_i}, \tag{1}$$

where r_k is the number of rewards earned on any particular option k, c_k is the number of choices made to that option, and the summations in the denominator are over all available options. In words, this expression states that the fraction of total choices that an animal allocates to an option will *match* the fraction of total rewards they earn on that option. This correspondence between reward and choice fractions is the central prediction of the matching law. The research presented in this article follows directly upon Herrnstein's, testifying to the continuing impact of his seminal work on animal choice.

Over the intervening decades, most studies of matching behavior have focused on the steady state—gathering data only after an animal's choice behavior equilibrates to any manipulation of reward contingencies. As shown by Davison and Baum (Baum & Davison, 2004; Davison & Baum, 2000) and by Gallistel and colleagues (Gallistel, Mark, King, & Latham, 2001; Mark & Gallistel, 1994), however, important mechanistic insights can be gained by examining the dynamics of the system as it operates in a state of flux. We therefore designed a dynamic foraging paradigm wherein animals' behavior must adapt to frequent changes in environmental conditions in order to gather rewards efficiently. As in the previous studies by Davison and Baum and by

Greg Corrado, Leo Sugrue, and William Newsome are at Howard Hughes Medical Institute and Stanford University School of Medicine. Sebastian Seung is at Howard Hughes Medical Institute and Massachusetts Institute of Technology.

Correspondence should be addressed to G. S. Corrado, Department of Neurobiology, Stanford University, D200 Fairchild Building, 299 Campus Drive West, Stanford, California 94309 (e-mail: gcorrado@alumni.princeton.edu).

doi: 10.1901/jeab.2005.23-05

Gallistel and colleagues, our primary goal is to gain insight into the mechanisms underlying matching behavior by studying the system as it operates near the limits of its adaptability.

We collected substantial behavioral data sets from 2 rhesus monkeys, some of the most flexible and tenacious reward harvesters in the animal kingdom (Southwick & Siddiqi, 1985). In earlier work with this data set, we employed a local formulation of the matching law, incorporating ''leaky'' integration of reward history, to model the animals' behavior (Sugrue, Corrado, & Newsome, 2004). This local matching rule captured the essential features of the data well and was more than adequate for the purposes of our earlier analysis, which focused on the interpretation of neurophysiological data. Our use of that model, however, was motivated primarily by its simplicity and formal similarity to the matching law, not by a principled exploration of possible alternatives.

We now take a very different approach. Rather than selecting a specific model and fitting it as well as possible to the data, we allow the data themselves to suggest the most appropriate model within a broad class of possibilities. Thus we aim to infer more directly the computations underlying choice behavior from the data themselves—to estimate rather than to fit. Our specific goal is to capture the dynamics of choice behavior within the broad framework of Linear-Nonlinear-Poisson (LNP) models (e.g., Chichilnisky, 2001). This class of models, which includes the leaky matching rule from our previous study, describes choice in terms of a feed-forward, three-stage process. However, rather than assume a specific functional form for each of these stages, here we reconstruct the function that best describes each stage directly from the raw data. To accomplish this, we employ an established sequential estimation procedure based on a general form of Wiener kernel analysis (Dayan & Abbott, 2001). Although our solution is constrained to lie within the LNP framework, this framework is far more general than our earlier casting of the data in the form of a leaky matching rule.

Linear systems analysis has been applied successfully to the analysis of reward-choice relations in elegant work by several research groups in the past. Linear techniques have been employed, for example, in studying the dynamics of extinction (Palya, Walter, Kessel, & Lucke, 1996, 2002), session-to-session changes in behavior under concurrent variable-interval (VI) reward schedules (Hunter & Davison, 1985), integration of reward effects over time (Horner, Staddon, & Lozano, 1997), and to establish a theoretical basis for the steady-state matching relation first enunciated by Herrnstein (McDowell, 1980; McDowell, Bass, & Kessel, 1983; McDowell & Kessell, 1979). Our use of the more general LNP framework both extends these methods and applies them in a new behavioral context.

As we will show, this approach ultimately recovers an LNP choice model that resembles our earlier leaky matching rule in several respects, but that also contains a number of key differences. These refinements include hyperbolic (as opposed to exponential) temporal weighting of past rewards, and differential (as opposed to fractional) comparisons of option value. We will demonstrate that this revised model successfully predicts single behavioral choices and independently generates realistic synthetic behavior. Finally, we will show how we can use this model to evaluate the optimality of our animals' behavioral strategy in terms of net rewards harvested.

EXPERIMENT

Figure 1A depicts our dynamic foraging task. In this task, two colored icons, or targets, appear on a computer screen, one red and one green. A computer monitors the animal's gaze continuously throughout the experiment, so when the monkey is instructed to make a choice it does so simply by moving its gaze from a central fixation cross to the desired target. The task progresses in rounds, or trials, in each of which the animal is free to make an eye movement to either (but not both) of the two targets. The eye movements are rewarded with drops of juice, delivered on a VI schedule.

On a VI schedule, rewards are delivered at random intervals, but at a constant overall rate and with a single constant magnitude. We implemented VI schedules using a Poisson process: at each point in time there is a constant probability that a reward will appear on a target. Once a reward is scheduled to appear on a target, say green, we say that that target is ''baited.'' That target remains baited

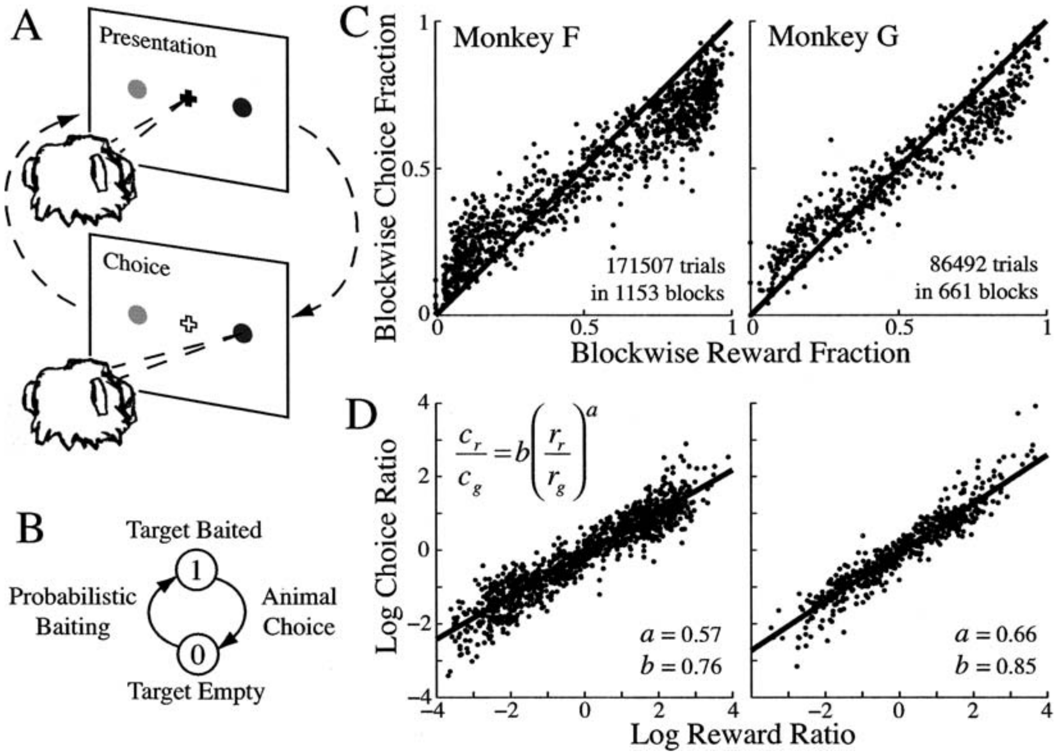


Fig. 1. (A) Schematic depiction of the foraging task. Subjects alternately view a *presentation* screen where they must hold their gaze on a central fixation marker (cross), and a *choice* screen where they are free to direct their gaze to either of the two colored targets, one red and one green. Rewards are delivered on dynamic VI schedules. (B) Schematic diagram of the process governing the state of a single target. Empty targets have a constant probability per unit time of being baited. Once baited, targets only become unbaited when the animal chooses said target and collects the reward. (C) Block-wise matching behavior for each of the 2 monkeys in our study. Each data point represents a block of trials on which the baiting probabilities for each target were held constant. Reward and choice fraction are shown here, and in all subsequent figures, relative to the red target (for the green target, the equivalent metrics are one minus the value for the red target). Thus the abscissa in 1C denotes the fraction of the total rewards in a particular block that were earned on the red target. (D) The same data from 1C is replotted as the log of the ratio of choices made to (or the rewards earned on) the two targets in each block. Blocks for which no rewards were earned on one or the other color are omitted to avoid data at +/− infinity. The data are fit by linear regression (solid line); the insets show the equation for the generalized matching law and the parameters of the fitted regression line.

until the next time the animal chooses green and collects the reward. Importantly, this process of target baiting and reward delivery (Figure 1B) occurs independently for the two colored targets.

The use of VI schedules is popular in classical matching studies because of their inherent compatibility with the behavior. When choosing between several options, each rewarded on a VI schedule, matching is very nearly the optimal policy for maximizing overall reward rate (Baum, 1981). Under VI schedules, therefore, the much-debated conflict between *matching* and *maximizing* as descriptions of behavior (e.g., Rachlin, Battalio, Kagel & Green, 1981; Vaughan & Herrnstein, 1987) largely disappears.

A second feature that our foraging task shares with many classical matching paradigms is the incorporation of a changeover delay (COD). The COD is a common technique for introducing a ‘‘cost,’’ in this case a temporal delay, to switching from one choice option to another (Shahan & Lattal, 1998). Consider, for example, a situation in which the animal chooses the red target for several trials and then switches to green. Under the COD, this first choice to green will not be rewarded even if that target is in fact baited—the baited reward will be delivered for a subsequent response to green only once the COD has expired. This delay is roughly analogous to the costs faced by animals in the wild that must abandon one foraging site and move to another before collecting additional rewards (Baum, 1982). The primary function of the COD is to deter the animal from switching rapidly between options, choosing each target only once. Without such a cost, an animal can gather rewards surprisingly efficiently by alternating between options—sometimes masking or even abolishing matching behavior (Stubbs, Pliskoff, & Reid, 1977). Thus, although graded matching behavior can be observed without the use of a COD (see, e.g., Lau & Glimcher, 2005), its incorporation shields the data from partial contamination with competing behavioral strategies based on alternation.

As noted above, our foraging task departs most significantly from classical matching paradigms in its use of dynamic reward schedules. In this regard, our task builds upon the dynamic designs introduced by Davison and Baum (Baum & Davison, 2004; Davison & Baum, 2000) and by Gallistel and colleagues (Gallistel et al., 2001; Mark & Gallistel, 1994). We expose animals to as many as a dozen changes in reward schedules over the span of a few hours. In practice, we held the sum of the reward-baiting probabilities on the two targets constant, so that on average a reward appeared on the targets at a combined rate of one every three choices. Thus only the *relative* reward-baiting probabilities were varied during our experiments.

METHOD

Subjects

Two adult male rhesus monkeys (*Macacca mulatta*) weighing 7 and 12 kg were used in this study. Prior to experimental use, each animal was prepared surgically with a head-holding device (Evarts, 1966) and a scleral search coil for monitoring eye position (Judge, Richmond, & Chu, 1980). Fluid intake was restricted outside of the experimental session but food was freely available. All surgical, behavioral, and animal care procedures complied with National Institutes of Health guidelines and were approved by the Stanford Institutional Animal Care and Use Committee.

Apparatus

Animals sat in a primate chair at a viewing distance of 57 cm from a color computer monitor. Their heads were positioned stably using the head-holding device, and eye position was monitored using a magnetic search coil apparatus (CNC Engineering, Seattle, WA). Behavioral control and data acquisition were managed by a PC-compatible computer running the QNX Software Systems (Ottawa, Canada) real-time operating system. The experimental paradigm was implemented in the NIH *Rex* programming environment (Hays, Richmond, & Optican, 1982). Visual stimuli were generated by a second PC-compatible computer and displayed using the Cambridge Research Systems *VSG* (Kent, UK) graphics card and accompanying software development tools. Liquid rewards were delivered to the animals via a gravity-fed juice tube placed near the animal's mouth, actuated by a computer-controlled solenoid valve. All subsequent data analysis and computer simulations were preformed on Apple Macintosh® computers in the Mathworks MATLAB (Natick, MA) programming environment.

Behavioral Task

Figure 1A is a schematic of the general structure of the foraging task. Animals alternately viewed a ''presentation'' screen where they were required to hold their gaze on a central fixation cross, and a ''choice'' screen where they were free to direct their gaze to either of two peripherally presented choice targets. These targets were of equal luminance but different color, one red and one green, and for any given experiment were presented at a pair of mirror symmetric locations in opposite hemifields.

Between successive occurrences of the presentation screen, the red and green targets were randomly assigned to the two prespecified locations, meaning that across trials each location was chosen with equal frequency irrespective of the baiting probabilities assigned to the two colors. On each trial, the presentation screen was first displayed for a variable delay period of 1 to 2 s; the appearance of the choice screen signaled the end of this delay period and cued the animal to indicate its choice with an eye movement to one of the two choice targets within a 1-s grace period. The animal was required to maintain its gaze on the chosen target for a further variable hold period of 300 to 600 ms. If the chosen target was baited at the time of the animal's choice, a fixed magnitude fruit juice reward (volume range 0.1 to 0.25 cc across experimental sessions) was delivered during this hold period. At the end of the hold period the presentation screen reappeared, cueing the animal to return its gaze to the fixation cross within a grace period of 1 s in order to trigger the onset of the next trial. Trials continued as long as the animal maintained its gaze within a 2° spatial window centered on the location of the fixation cross or chosen target. When the animal's gaze deviated outside of this window inappropriately, the trial was terminated and a 2 to 4 s timeout period elapsed before the presentation screen reappeared. This temporal structure encouraged the animal to execute long sequences of trials mimicking continuous foraging behavior in a trial-based setting.

Target color was the cue to the rate of reward, specifying the probability per unit time that an empty target would be baited with a reward. The sum of the reward-baiting probabilities on the two colors was held constant at a rate of 0.12 rewards per second, resulting in a combined reward-baiting probability of approximately 0.3 rewards per trial. Meanwhile, the relative reward-baiting probabilities on the two colors were changed without signal between blocks of trials that varied in length between 50 and 300 trials. On each block, the relative baiting probabilities on the two colors was chosen unpredictably from a subset of the ratios: 1:8, 1:6, 1:3, 1:2, 1:1, 2:1, 3:1, 6:1, and 8:1 (though not every daily experiment used all ratios).

The clock that implemented the reward schedule on each color ran whenever that color was unbaited and the presentation or choice screens were displayed; during timeout periods the scheduling clock was paused and the states of the targets (baited or unbaited) were preserved until the presentation screen reappeared. Once baited with a reward, a particular target color became unbaited only when the animal choose that color and collected the reward (Figure 1B), at which point the scheduling clock for that color began running again. A COD was in effect continuously during data collection. Enforcing the COD meant that reward delivery was withheld for choices of baited targets that constituted a switch between target colors; on such trials, the chosen target remained baited and its reward was delivered normally upon a second successive choice of the same color.

Training

Each animal was trained over a period of 4 to 5 months during daily sessions that lasted 3 to 4 hr. The endpoint for training was the production of reliable matching behavior of the type depicted in Figure 1C and D. Training occurred in sequential phases that progressed to this final goal, with the time to accomplish each phase varying somewhat between monkeys.

Phase I: Single target task. In the initial phase of training, animals learned to fixate a centrally appearing fixation cross and to make accurate saccadic eye movements back and forth between this fixation cross and a single purple colored target presented elsewhere on the display. The timing of these movements was cued by simultaneous increases or decreases in the luminance of the fixation cross or target. Initially, all correctly executed movements were rewarded with the delivery of a fixed volume of juice with a probability of 1.0. Once the animal was accustomed to this sequence of movements, the luminance of the peripheral target was held constant so that changes in the luminance of the fixation cross alone cued the timing of movement execution, and rewards for movements to the fixation cross were eliminated. If the animal failed to make an appropriate eye movement, either to the target or back to the fixation cross, the task entered a 10 to 20 s timeout period. This timeout served as a punishment and was used to encourage the animal to return to the fixation cross when cued and begin the next trial. In this manner, animals learned to link successive trials into uninterrupted series during which the timing of the animal's behavior was under continuous experimental control. Once the animal was reliably performing series of 5 to 10 trials, the probability of reward for a correctly executed movement to a target was gradually reduced from an initial probability of 1.0 to a goal level of approximately 0.3 (corresponding to a baiting probability of 0.12 rewards per second).

Phase II: Choice task. After the animal was efficiently performing the single target task with the same overall baiting probability (0.12 rewards per second) and timing of events that would be used in the final choice task, we introduced the choice situation in which two differently colored targets were presented simultaneously on each trial. These red and green targets were otherwise identical to the single purple target used in the earlier task, as was the timing of behavioral events. Between successive trials, red and green targets were randomly assigned to the available spatial locations. During initial training, a large number of spatial locations were used in each session to discourage the development of spatial biases; later in training (and during the collection of most experimental data) only two locations equidistant from the central fixation cross were used.

Phase III: Extreme schedules. Initially, 100% of available rewards (at an overall reward rate of 0.12 rewards per second) were scheduled on one or other target color, and the animal's task was to identify the rewarded color and choose this color irrespective of where it was presented. Within and across daily sessions, the identity of the rewarded color was reversed while monitoring for corresponding reversals in the animal's choice behavior. During the first few of these sessions, an increase in the luminance of the rewarded target provided the animal with an additional cue to the identity of the rewarded color and aided in reversal of choice behavior. This luminance cue was quickly eliminated, and over subsequent sessions the frequency of unsignaled reversals was increased until the animal routinely encountered more than a dozen unsignaled reversals within a single session. To prevent the animal from anticipating their occurrence, the time between reversals was varied widely across a range of 50 to 300 trials.

Phase IV: Equal schedules. Once animals were reliably reversing their choice behavior in the setting of extreme reward schedules, blocks of trials were introduced in which the relative baiting probabilities on the two colors were varied between these extremes. Animals were initially exposed to blocks in which the baiting probabilities on the two colors were equal. At first, animals continued to show exclusive preference for one or the other target color during these equal probability blocks, a strategy that reduced the animal's reward rate by 50%. With increased experience, however, animals began to distribute their choices between the targets, eventually allocating 50% of their responses to each color as predicted by the matching law.

At this stage of training, animals began to show a strong tendency simply to alternate their choices between colors on successive trials. As mentioned above, the emergence and efficacy of alternating strategies in the context of concurrent VI VI schedules is well documented (Baum, 1974; Stubbs et al., 1977). This tendency to alternate was rapidly eliminated with the introduction of a COD, as described earlier, for the rest of the experiment. After implementing the COD, each animal's stay durations—the number of consecutive choices of a particular color before switching to the other color—assumed an exponential distribution consistent with a probability of switching that was constant on every trial. Such stochastic switching is characteristic of matching behavior (Gallistel & Gibbon, 2000; Heyman, 1979).

Phase V: Intermediate schedules. Equal probability blocks were initially interleaved with extreme blocks in which over 90% of rewards were assigned to one or the other color. Once animals were reliably changing their allocation of behavior between blocks of extreme and equal baiting probabilities, blocks that employed intermediate ratios of baiting probabilities were introduced. Training continued until animals reliably matched their choice and reward fractions across blocks of variable length on which the relative baiting probabilities on the two colors were chosen unpredictably from the set: 1:8, 1:6, 1:3, 1:2, 1:1, 2:1, 3:1, 6:1, and 8:1.

Phase VI: Experimentation. Once the animals showed reliable matching behavior on randomly interleaved schedules, the study entered its experimental phase. The conditions of experimentation were identical to those in the final stage of training. All of the data presented in this article were collected during this period.

RESULTS

Rapid Adaptation and Locally Driven Foraging Behavior

Our animals generated good matching behavior in a highly dynamic foraging environment. Figure 1C (following traditional depictions of matching behavior) plots the proportion of choices each monkey made to that target (*choice fraction*) as a function of the proportion of rewards each animal received from a particular target (*reward fraction*) over their entire experimental histories. Each point represents behavioral data collected during a single block of 50 to 300 trials where the reward-baiting probabilities of the two targets were held constant.

According to the strict matching law (Herrnstein 1961, 1970), these points should lie along the unity line, that is, the fraction of choices made to the red target should exactly match the fraction of rewards earned on red. Although the observed behavior conforms to the prediction generally, both animals show a similar tendency toward modest undermatching: the data points are distributed along a line with a slope somewhat shallower than unity. This tendency toward undermatching is a well-documented deviation from Herrnstein's original law, common in many behavioral experiments (Baum, 1979).

Such deviations are typically quantified in terms of the generalized matching law introduced by Baum (1974). Baum recognized that Herrnstein's original formulation (Equation 1) can be expressed equivalently in terms of ratios, rather than fractions,

$$\frac{c_1}{c_2} = \frac{r_1}{r_2}. \tag{2}$$

This ratio expression is easily generalized to incorporate parameters that characterize the typical deviations from pure matching observed in real data:

$$\frac{c_1}{c_2} = b\left(\frac{r_1}{r_2}\right)^a. \tag{3}$$

In this generalized formulation, the parameter a quantifies the sensitivity of choice to changes in the ratio of rewards, and the parameter b captures any bias toward one or other response option independent of reward. Logarithmically transforming this equation,

$$\log\left(\frac{c_1}{c_2}\right) = \log(b) + a\log\left(\frac{r_1}{r_2}\right), \tag{4}$$

suggests immediately that these parameters can be estimated by plotting the data in a log-ratio space and measuring the slope and intercept of a regression line (Figure 1D).

In a review of experiments that studied matching behavior under steady-state conditions, Baum (1979) found sensitivities that were typically in the range of 0.8 to 0.9. The sensitivities that we obtained, 0.57 and 0.66 for Monkeys F and G, respectively, are lower than those seen in steady-state experiments, but are in good agreement with those reported in other studies that have examined matching under dynamic conditions (e.g., Davison & Baum, 2000). Lower sensitivities under dynamic conditions should come as no surprise given that such conditions violate the basic assumption that behavior has equilibrated at steady state. Indeed, any analysis that averages data across an entire session/block is ill suited to experiments that incorporate frequent changes in reward contingencies. Instead, dynamic conditions require analytic techniques that capture the evolution of choice allocation as reward contingencies change within a single experimental session.

We now address these dynamics. Figure 2A plots the time-course of typical behavior from Monkey F across a single experimental session. The thin black line shows the reward-baiting probability assigned by the experimenter to the red target (expressed as a fraction of the total reward probability) as the computer steps through the blocks of the experiment. In the second block, for example, the probability of a red target being baited was twice the probability of the green target being baited, thus the fractional baiting probability is 2/3 in favor of red. The dashed and thick lines depict the monkey's reward fraction and choice fraction, respectively (again relative to the red target), for the first, middle, and last third of each block.

The generally similar time-course of the three traces show that the animal's behavior is effectively influenced by our experimental manipulations. The similarity of the thin and dashed lines indicates that altering the baiting probability impacts the animal's experienced reward fraction. This correspondence is ap-

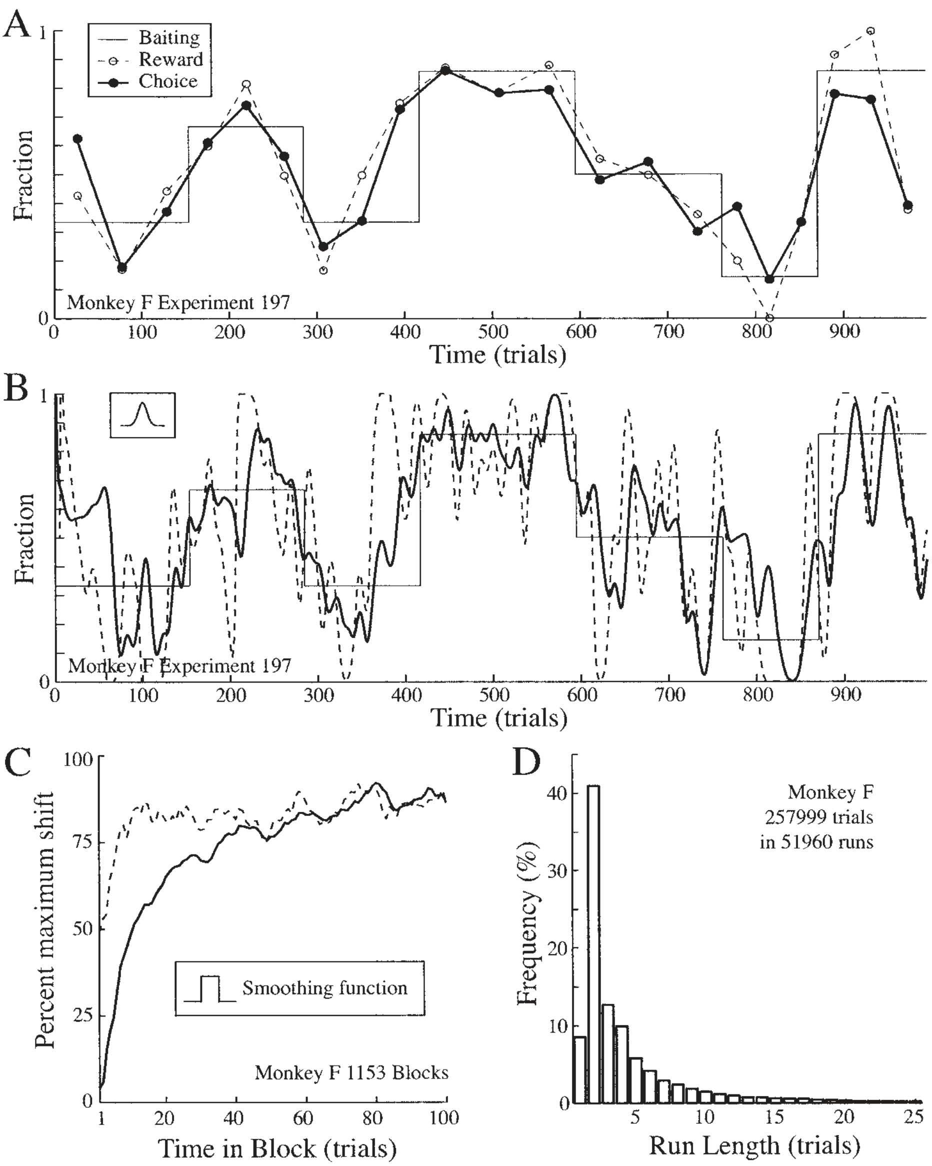


Fig. 2. (A) Time-course of reward and choice fractions for a single experiment. The thin line shows the fraction of the total baiting probability assigned by the experimenter to the red target. The dashed line shows the resulting experienced reward fraction for the red target, calculated for the first, middle, and last third of each block. The thick line shows the animal's choice fraction for the red target over the same period. (B) High temporal resolution view of reward and choice fraction time-courses. The data are the same as in A, but with reward and choice fractions computed locally using a Gaussian filter (inset), rather than chunked by thirds of a block. (C) Behavioral response to block transitions. The dashed and solid curves plot the average time-course of adaptation of reward and choice fractions after a block transition, in normalized units where the fractional baiting probability on the previous block is 0% and the fractional baiting probability on the new block is 100%. Reward and choice fractions are computed using a box filter (inset) before averaging across blocks. (D) Distribution of run lengths for Monkey F across all experiments. Each bin shows the relative frequency of choosing a target exactly *n* consecutive times before returning to the other option.

proximate rather than exact, both because of the stochastic nature of the VI baiting schedules and because the animal's own choices affect its experience of rewards. (If the animal were to choose only red, for example, the experienced reward fraction in Figure 2A would be unity irrespective of the programmed baiting probabilities.) The similarity between the dashed and thick curves, meanwhile, confirms that the monkey's allocation of choices conforms to the matching law even at the timescale of these shorter epochs. These correspondences seem to suggest that baiting fraction (thin) influences reward fraction (dashed), which in turn guides choice (thick).

We found that the animals' behavior adapted quickly to changes in the environment at block boundaries. We analyze this process of adaptation in Figure 2C, where we plot the shift in reward fraction and choice fraction as a function of time since the onset of the block transition, averaged across 1,153 blocks for Monkey F. The shifts are shown in percentage change relative to the new blockwise baiting probability, thus a 0% shift in reward fraction would indicate that it matched the fractional baiting probability on the previous block, whereas a 100% shift would indicate that it matched the fractional baiting probability on the new block. Fractions are calculated over five trials (smoothing function inset), normalized to the percentage of the maximum expected shift, and then averaged across blocks. Following the convention in panel A, the reward and choice are shown by dashed and thick lines, respectively.[1] Recall that a shift in observed rewards is the animal's only cue that the environment has changed, and thus sets an upper limit on the rate at which choice behavior can adjust. Only 10 trials after the block transition, when the shift in reward fraction is just becoming stable (dashed line), the animal has already made the bulk of the choice fraction adjustment (thick line). After 40 trials, the animal's behavior has adapted completely to the new environment.

Figure 2B demonstrates more precisely the temporal fidelity of the relation between reward and choice. This panel replots the data shown in 2A, but now as instantaneous estimates of reward and choice fraction rather than as chunked averages. To compute this more local measure, we begin by smoothing the binary record of rewards or choices with the Gaussian function shown in the inset (standard deviation five trials). We then compute the reward and choice fractions using these smoothed signals, allowing a much closer look at the variability and structure of the behavior.

This closer examination reveals two interesting features that were previously hidden. First, the stochastic fluctuations in reward fraction (dashed line) *within* each block rival in magnitude those induced by our experimental manipulation of baiting probabilities *between* blocks. Secondly, the animal's behavior (thick line) seems to follow these stochastic fluctuations in reward fraction as tightly as it does the changes at block boundaries. Essentially, the animals appear to track the noise in the reward rates as aggressively as they track ''real'' changes in the environment. If this is indeed the case, then adherence to the matching law is exquisitely local in time—choice patterns track reward patterns not at the time scale of blocks, but moment by moment. (As a corollary, it would seem unlikely that the animal is able to distinguish stochastic fluctuations from actual changes in the state of the world. We test this hypothesis below by determining how well a model based on this principle mimics the animals' actual behavior.)

Figure 2D illustrates another important feature of animal behavior in this task: within the overall constraint of matching behavior, choices appear to be stochastic. The figure is a frequency histogram of run length—consecutive choices to a target of a single color—averaged across all behavioral sessions for Monkey F. (The frequency histogram for Monkey G is similar, and is illustrated in Figure 7B.) The distribution exhibits an approximately compound exponential form (verified quantitatively in the section entitled *Model Validation* below), consistent with probabilistic choice behavior and conforming to Bernoulli statistics. The notable exception to the exponential form is that runs of length

[1] In Figure 2C, the average reward fraction asymptotes below 100% of the expected shift because of bias introduced by asymmetry in the distribution of observed reward fractions. As illustrated in the instantaneous plots in Figure 2B, there are substantial deviations of the observed reward fraction from programmed baiting fraction. These excursions are not distributed symmetrically about the mean, but are heavily skewed because of hard barriers to fractional reward, which obviously can never be less than 0 or greater than 1. For example, in all the blocks shown in Figure 2A and B except for Block 5, there is more room to deviate *below* 100% expected shift than there is to deviate *above* it. This asymmetry induces a bias, which tends to pull the mean below 100%. Consistent with this analysis, plots similar to Figure 2C, constructed with medians as opposed to means, actually asymptote slightly above 100% of expected shift, rather than below.

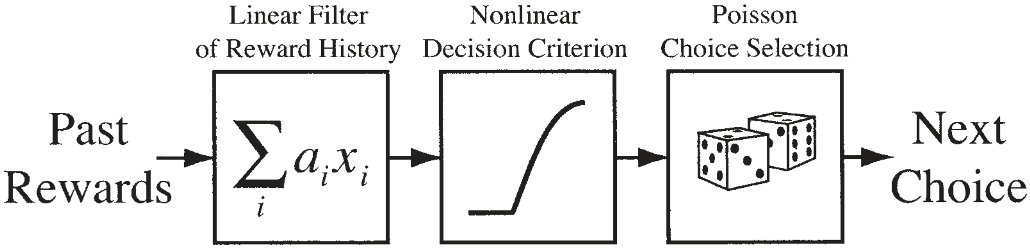


Fig. 3. Diagram of a generic LNP model for choice. Past rewards enter at the left, coded as a binary stream. A linear filter weights the rewards based on their distance in the past. The resulting scalar is mapped to probability of choice by a low-dimensional nonlinear function. That probability of choice is then used to drive an inhomogeneous Poisson process, which in turn renders the ultimate binary choice.

one are massively under-represented. This, of course, is the influence of the COD, which ensures that a single isolated choice to one color is never rewarded. As a result, after a single choice to a new color, Monkeys F and G made a second choice to that color 92% and 96% of the time, respectively.

This distribution of run lengths is particularly germane to the historical characterization of matching behavior as exhibiting a constant leaving rate (Heyman, 1979), which is to say that animals abandon the option that they are currently choosing with a constant probability per unit time. In this scenario, the run length should in fact be distributed exponentially, or in the discrete approximation, geometrically. In our dynamic environment, the leaving rate is not in fact constant, but instead changes as the reward contingencies change from block to block. In this case, run length should be distributed as a mixture of exponentials of various time constants, resulting in a somewhat heavier tailed distribution than a pure single exponential. Intuitively, when applied to a dynamic environment, the constant leaving rate hypothesis implies that the animal flips a coin on each trial to decide which option to choose, with the weight of that coin varying parametrically with recent reward history.

A Framework for Behavioral Strategy Estimation

The preceding analysis suggests that whereas aggregate behavior abides by the matching law, the animals' decisions are guided by a mechanism that is inherently local in time. There are, however, many local behavioral mechanisms that could give rise to the global property of matching (e.g., Sugrue et al., 2004; Vaughan, 1981). To study the behavioral mechanism that animals employ, we must work within some framework for describing possible foraging strategies—one general enough to capture the observed behavior, yet constrained enough that inference of the particular underlying strategy from data is tractable. We have chosen to work within the class of Linear-Nonlinear-Probabilistic (LNP) models, of which our earlier leaky matching rule is an example. We will validate our choice of framework and the associated estimation procedure post hoc, through both predictive and generative testing of the recovered behavioral model.

The LNP framework, diagrammed in Figure 3, consists of three serial feed-forward computational stages. The generic form of an LNP model is as follows: In the first stage (L), a linear operator projects the high dimensional input (reward history) onto a lower dimensional output. In our case, this will prove to be a simple linear filter of recent reward history that captures the influence of reward on choice with a single scalar parameter. In the second stage (N), a static nonlinear function remaps the output of the L-stage onto a single decision probability. In the final stage (P), a point-wise independent random variable, or Poisson process, draws an outcome with this probability, thus rendering the ultimate binary choice.

One aspect of choice behavior that we do not expect the LNP framework to capture is the influence of the COD. If the animal switches its choice of color on trial t, then its choice on trial $t+1$ is a foregone conclusion—as stipulated by the COD, it must again choose the color it switched to lest its previous response be wasted. Thus the COD-related trials amount to simple duplications of the previous response—an independent strategy that preempts the matching strategy occasionally (20% of all choices for Monkey F and 19% for Monkey G). Because matching behavior is the focus of our analysis, we direct our modeling efforts toward explaining *free choices*, those that are guided by the underlying foraging strategy and unfettered by the particular form of our COD penalty.

Estimation of Behavioral Strategy

A strength of the LNP framework is that the best of a family of models can be estimated from data sets of modest size (e.g., Horwitz, Chichilnisky, & Albright, 2005). We have estimated the serial stages of the LNP process sequentially so as to eliminate the computational challenge of optimizing all three stages simultaneously. Under certain assumptions, this sequential optimization can produce completely unbiased estimates, and the flaws introduced by modest violations of these assumptions are small (Bussgang, 1975, Simoncelli, Paninski, Pillow, & Schwartz, 2004). We show in the Appendix that this sequential estimation procedure is both stable and sound for our specific application, despite not being provably unbiased (see Appendix Figure A1).

Our first challenge is to estimate the linear (L) stage of our model. The usefulness of linear operators in describing behavior was established by Palya et al. (1996, 2002), and in some sense this is the most critical aspect of our modeling effort. It is the L-stage that looks back at the actual history of rewards on the last N trials and distills the influence of all of those events into a single number. This scalar captures the relative value of choosing one or the other option given recent events, as evaluated by the particular form of the L-stage—for convenience we will refer to this quantity as the model's *scalar value metric.* Because we only consider linear functions for our L-stage, our scalar value metric is computed by multiplying a reward event that occurred i trials ago by a weighting coefficient $k(i)$ and summing the results for all values of i. In its most general form this can be written:

$$L[r_r(t),\, r_g(t)] = \sum_{i=1}^{N} k_r(i) r_r(t-i) + \sum_{i=1}^{N} k_g(i) r_g(t-i), \qquad (5)$$

where $r_r(t)$ and $r_g(t)$ are binary vectors denoting rewards obtained on the red and green targets on trial t. (It is important to note that to preserve the meaning of t as a temporal index, *all* trials appear in *both* reward histories regardless of which color was chosen on the trial. Because on each trial t, the reward vector $r_q(t)$ is 1 if and only if color q was chosen and rewarded, this implies that $r_q(t)$ will be 0 either if color q was chosen on trial t and not rewarded, or simply because color q was not chosen on that trial.)

If we assume that the weighting of red and green rewards is similar, and that their impact on choice is equal and opposite, we can rewrite our L-stage operator more simply

$$L[r(t)] = \sum_{i=1}^{N} k(i) r(t-i), \qquad (6)$$

where we define a composite reward history

$$r(t) \equiv r_r(t) - r_g(t). \qquad (7)$$

In this format, it is easy to recognize that different formulations of the weighting coefficients $k(i)$ will implement different linear operations on the recent reward history. If, for example, we chose $k(i) = 1/10$ for $i = 1$ though 10 and $k(i) = 0$ for $i > 10$, then our linear operator would compute a scalar value metric that is a signed average of the rewards received over the last 10 trials. If instead we chose $k(1) = 1$, $k(2) = -1$, $k(i) = 0$ for $i > 2$, then our linear operator would report the signed difference of the rewards received on the last two trials. Thus it is the set of weighting coefficients k, sometimes called the kernel, that describes the nature of the filter.

Our task, however, is to find the N coefficients of the kernel $k(i)$ that best relate recent reward history to subsequent choice for the strategy our animals actually employ. To that end, we will try to find the kernel that gets the output of our L-stage as close as possible to the final output we want our model to capture, the actual choice made by the animal. More formally, we would like to find a set of $k(i)$ that minimizes the sum of squared errors between the output of the L-stage for each trial and the choices made by the animal on that trial,

$$E = \sum_{t \notin \mathrm{COD}} \{L[r(t)] - c(t)\}^2, \qquad (8)$$

where choice has been coded as a composite history much like reward, $c(t) = c_r(t) - c_g(t)$, based on separate binary choice vectors $c_q(t)$, which are 1 if color q was chosen on trial t and 0 if it was not. (Of course, because one or the other color is chosen on every trial, $c(t)$ is comprised entirely of +1 and −1 entries, whereas $r(t)$ is mostly zeros simply because most trials are unrewarded.) As discussed earlier, we are only interested in predicting behavior on free choices, and for this reason, we accumulate error in this sum only over trials not governed by the COD.

Fortunately, we do not have to invent a method for solving this optimization problem. The problem of finding a discrete linear filter, defined by $k(i)$, which transforms an input time series $r(t)$ to most closely match a desired output time series $c(t)$, is a very common problem in linear systems analysis and digital signal processing. Norbert Wiener famously proved that the solution to this optimization could be found analytically by solving the Wiener-Hopf equation:

$$C_{xx}\vec{k} = \vec{C}_{xy}. \tag{9}$$

The essence of Wiener's proof was to demonstrate that the kernel k that minimizes the sum of squared errors between the time series if the input $x(t)$ and the output $y(t)$, is the same k that solves this much simpler matrix equation. C_{xy} in this context is the cross-covariance series between the input and the output, and C_{xx} is the auto-covariance matrix of the input. (This method is extremely closely related to cross-correlation and reverse-correlation methods common in behavioral and neurophysiological studies, which some readers may be more familiar with. Please see Kay, 1993, for an introduction to this and related analyses.)

To apply Wiener kernel analysis to our problem, we must make a few adjustments. First, because we only wish to allow our filter access to *past* rewards in predicting the current choice, we must additionally constrain our analysis to recover only purely causal filters. We enforce causality by restricting our covariance calculations to positive lags (those that correlate the present to the past) and discarding data at zero lag (that correlates present choice to the resulting reward) or negative lags (that correlate present choice to future reward).

With this in mind, we might consider using the reward time series $r(t)$ as the input and the choice time series $c(t)$ as the output. We could then use the Wiener-Hopf equations to directly compute the kernel k by inverting the auto-covariance of $r(t)$,

$$\vec{k} = C_{rr}^{-1}\vec{C}_{rc}. \tag{10}$$

However, as written, this would provide the best linear filter accumulating error over all trials, including the COD-related trials. To remove the contribution of COD trials to the cross-covariance, C_{rc}, we must make the additional modification of defining an alternative free-choice time series to use as the target output,

$$\hat{c}(t) = \begin{cases} 0 \text{ if trial } t-1 \text{ was a switch} \\ +1 \text{ if free choice to red on trial } t \\ -1 \text{ if free choice to green on trial } t \end{cases}$$

Because this free-choice vector has zeros wherever the COD governed the choice, these trials contribute nothing to the cross covariance, provided that the mean of the choice vector $c(t)$ is zero. The same effect could be achieved by simply leaving out nonfree choice trials when constructing the cross-covariance, but the former procedure is far more computationally efficient.

We can now write an equation for our estimate of the best linear filter relating rewards to choice:

$$\vec{k} = C_{rr}^{-1}\vec{C}_{r\hat{c}}. \tag{11}$$

Because the overall amplitude of k is immaterial, we can additionally normalize k so that

$$\sum_{i=1}^{N} k(i) = 1, \tag{12}$$

without reducing the representational power of our LNP model. (The reason scaling factors on k do not impact the generality of the model is that any useful multiplicative factors on the scalar value metric can be subsumed entirely by the nonlinear N-stage of our model, and so we can remove them at this stage and focus just on the relative size of the weights $k(i)$.)

The data points in Figure 4A illustrate the Wiener kernel weighting coefficients recovered in this manner for each of the 2 monkeys in our study averaged across their entire data sets. To obtain these average kernels, we computed C_{rr} and $C_{r\hat{c}}$ for each day's experiment and averaged them across experiments, weighting each by the number of free choices the animal made on that day. Kernel recovery was then based on these average covariances. The kernels reconstructed for the 2 animals are quite similar in overall shape, exhibiting a super-exponential, or heavy-tailed, appearance. In contrast to the single exponential profile assumed in Sugrue et al. (2004), these reconstructions reveal the L-stage kernel is better described as a mixture, or weighted sum, of two exponentials,

$$k(i) = a\frac{1}{n_1}e^{-\frac{i}{\tau_1}} + (1-a)\frac{1}{n_2}e^{-\frac{i}{\tau_2}}, \qquad (13)$$

where n_a are normalization constants such that,

$$\sum_{i=1}^{N}\left(\frac{1}{n_a}e^{-\frac{i}{\tau_a}}\right) = 1. \qquad (14)$$

The black traces show the best double-exponential parameterization of the recovered kernels. We acquired these fits by minimizing the mean squared error between the raw kernel weights we recover and the parameterized double-exponential. The insets display the fit parameters for each monkey. The short exponential seems to reflect rewards accrued over the past one to two trials, whereas the long exponential extends 10 to 20 trials into the past. Both animals put more weight on the longer timescale exponential, τ_2. It is important to contrast our current estimation approach with the approach of fitting the data to an assumed functional form (as was done in our earlier work). Here we use a double-exponential only to describe the shape of the linear filter that was obtained from an unprejudiced analysis of the data. We made no a priori assumptions about the filter's shape but instead recovered that shape from direct Wiener kernel analysis of the behavioral data.

To summarize, our L-stage takes as its input the reward history on each of the two color targets, $r_r(t)$ and $r_g(t)$, and outputs a scalar value metric for each trial t. The scalar value metric is computed by the linear filter, $L[r(t)]$ given in Equation 6, based on a double exponential kernel $k(i)$ described by Equation 13. For convenience we will refer to this scalar value metric as *differential value.*

With this result in hand, we move to the task of estimating the N-stage of our LNP model. The role of the N-stage is to remap the output of our L-stage, the scalar value metric differential value, onto a probability appropriate for driving the P-stage. Thus estimation of the N-stage amounts to nothing more than finding the function that relates the differential value computed on trial t to the animal's instantaneous probability of choosing red or green on that trial. To map this relation, we bin trials according their differential value and simply compute the frequency of red choices the animal made on that subset of responses. Because observed frequencies yield maximum likelihood estimates of underlying probabilities, this direct approach produces a maximum likelihood estimate of the N-stage.

The data points in Figure 4B depict the recovered N-stages for both monkeys, again averaged over the entire data set for each. We grouped the choice data for each monkey into 30 equally populated bins based on the output of the best-fitting L-stage for that animal. Again, we omitted choices governed by the COD so that only free choices contribute to the results. The 95% confidence intervals on these probability of choice estimates, based on binomial statistics, are smaller than the plotted points. The recovered nonlinearity appears to be sigmoidal. We therefore fit cumulative normal functions (black traces) to the estimated N-stage nonlinearities. The equation for a cumulative normal used for these fits follows the standard form:

$$\Phi(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x, \qquad (15)$$

μ being the bias, or mean, of the underlying normal distribution; σ, the variance, or width. The cumulative normal fits describe the N-stage nonlinearity reasonably well, though the shape of Monkey G's nonlinearity shows clear signs of additional complexity for small differential values. The insets in the figures provide exact parameters of the fits. Note again that we use curve fitting only as a secondary tool to compactly describe a transform that was recovered directly by maximum likelihood estimation.

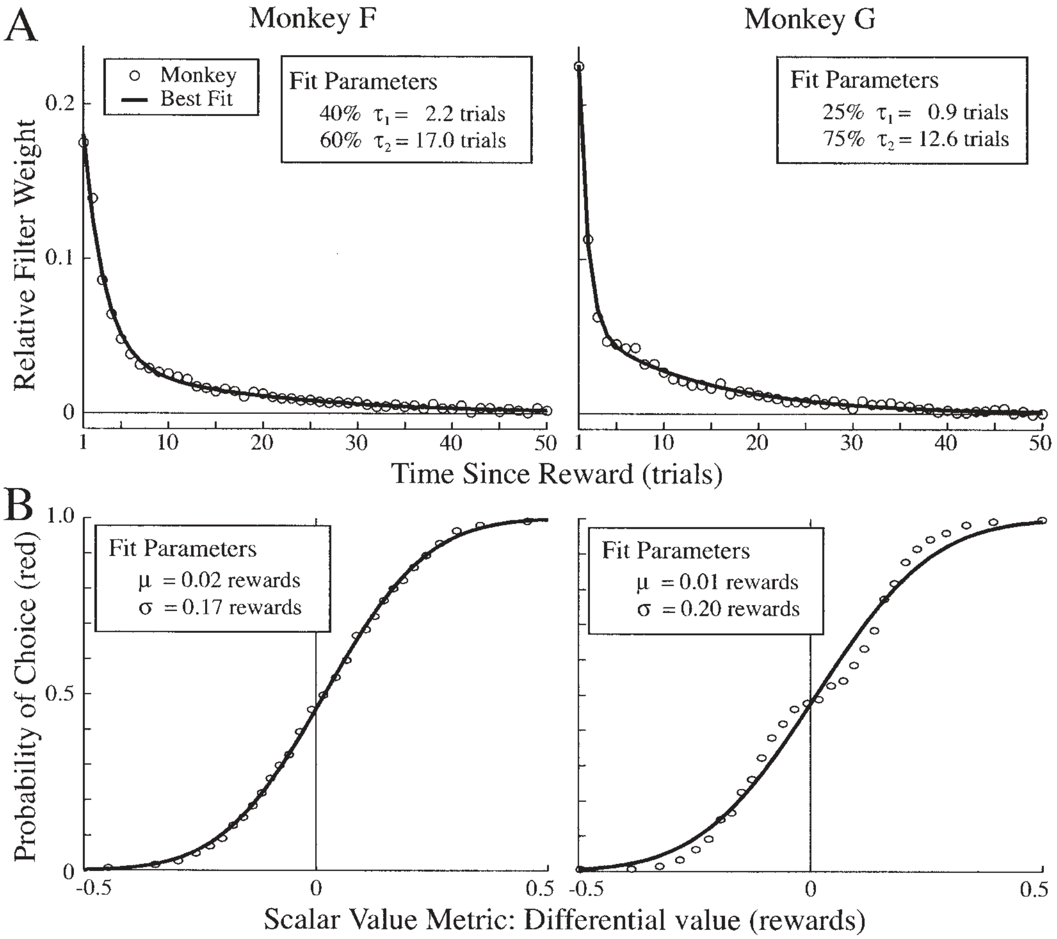


Fig. 4. (A) Filters estimated for the L-stage of our LNP model. The points indicate the raw filter weights recovered by Wiener kernel analysis for each monkey. These weights show the relative influence of rewards earned on each of the last 50 trials on the animal's subsequent choice. The filters are normalized so that the recovered weights sum to 1.0. The solid line shows the best double exponential fit to these raw filter weights; the inset shows the parameters that describe this double exponential. (B) Nonlinear decision criterion estimated for the N-stage of our LNP model. For each monkey, we show the relation between the scalar value metric, differential value, as computed by the L-stage filters shown in A, and the animals' ultimate probability of choice. The data points are equally populated bins showing the fraction of choices made to the red target for trials when the filtered reward stream held a particular value. Only free choices are included in computing these probabilities. The solid line shows the best-fitting cumulative Gaussian. Parameters describing this cumulative Gaussian are shown in the inset.

Figure 5 diagrams the final LNP model relating reward history to future choice. The inputs are binary reward histories for each color (left) that code whether or not a reward was received from that color on each trial in the past. Note that these histories contain a zero for any trial on which the color in question did not deliver a reward, whether or not that color was chosen. In the L-stage, these reward histories are convolved with the heavy-tailed linear filter that we recovered by Wiener kernel analysis and parameterized as a double-exponential. The difference of the outputs of these linear filters yields an intermediate scalar value metric that we term differential value. This scalar value metric operationalizes the relative value of the two choice options on the current trial. (Here we describe the difference operation as following filtering, but it is exactly equivalent to imagine differencing reward histories prior to filtering.) In the N-stage, a sigmoidal operator that we recovered by maximum-likelihood estimation and parameterized as a cumulative Gaussian remaps

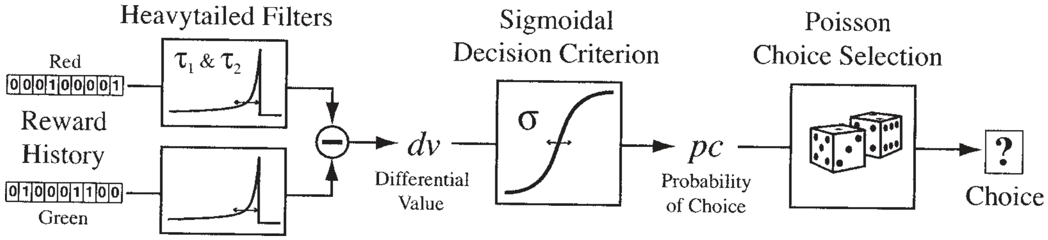


Fig. 5. Diagram of our final parameterized LNP model. Past rewards enter at the left, separately coded as a binary stream for each color. Rewards are coded as zeros if the animal did not choose the target on that trial, or if it chose the target but was not rewarded. Two identical double exponential filters weight these rewards based on their distance in the past. The difference of the output of these two filters is an intermediate scalar value metric we term differential value. Differential value is mapped to probability of choice by a sigmoidal decision function, parameterized as a cumulative Gaussian. That probability of choice is then used to drive an inhomogeneous Poisson process, which renders the ultimate binary choice.

differential value (range, −1 to +1) onto the scalar probability of choice (range, 0 to +1), which dictates the probability of choosing red on the current trial. In the P-stage, a binary random variable driven by the probability of choice renders the ultimate selection of the next choice (red or green). The behavior of the model is completely described by five parameters: τ_1 and τ_2, the short and long timescale components of the linear filter, *a*, the mixing coefficient for these two components, σ, the width of the sigmoidal nonlinearity, and μ, the residual bias. We will explore the effects of varying these parameters on choice behavior in a later section.

Model Validation

The estimation procedure discussed in the previous section would return a model even if the underlying data were fundamentally not well described by the LNP framework. Thus we need a way to validate our model independently, by comparing its output to real animal choice behavior. This requires us to demonstrate the model's sufficiency in two distinct but equally important senses.

First, we demonstrate our model's *predictive* sufficiency. Assessing predictive sufficiency amounts to asking the question: Given the animal's reward history up until time *t*, how well can we predict its choice at time $t + 1$? Second, we demonstrate the model's *generative* sufficiency. For this we ask the question: Thrust into the same foraging environment for extended blocks of trials, does the model behave in a manner that is qualitatively and quantitatively similar to the real animal?

Predictive Sufficiency

Figure 6 addresses the question of predictive sufficiency. Figure 6A (dashed line) shows the same choice data from Monkey F as Figure 2B; we compare this behavior with the predictions of our LNP model for Monkey F (solid line). So that there is no possibility of unfair advantage, the model parameters used were reestimated omitting the data from this experiment before predictions were made. (In practice, this resulted in a less than 1% deviation in the model parameters in Figure 4.) To generate predictions, we use the animal's actual reward history over the previous hundred trials as the model input. Based upon this history, on every trial, the model computes the animal's probability of choosing each target and renders a binary choice with that probability. This series of choices produced by the model was smoothed with the same kernel used to smooth the monkey choice data, inset in Figure 2B.

If the model is predicting choice effectively, the solid and dashed traces in Figure 6A should be very similar. Looking at this one example experiment, the LNP model appears to do well in predicting the animal's actions. Note, for example, that the solid prediction trace in 6A is much more tightly correlated with choice (both in temporal lag and magnitude) than the dashed reward trace in 2B. This degree of correspondence between predicted and observed behavior is typical across experiments and across animals.

Figure 6B quantifies the predictive sufficiency of our LNP model averaged across the entire data set for each animal. On each trial, we calculate the probability of choice predicted by our model (i.e., the output of the N-stage, before the P-stage has rendered a binary choice) using the animals' actual reward history as the input to our LNP model. Generalizing our approach above, we generate these predictions using a modified form of *leave-one-out cross validation*: leaving out each experiment in turn, reestimating model parameters based on that reduced data set, and then making predictions for the one omitted experiment. Based on these predicted probabilities of choosing the red target, we sorted all free choice trials into 30 equally populated bins and calculated the observed probability of choosing the red target for the trials in each bin. The 95% confidence intervals for these probability estimates are smaller than the plotted points. Were our LNP model a perfect predictor of the animal's probability of choice, the data in Figure 6B would lay exactly on the unity diagonal. We can see that the model does well at predicting average probability of choice across its entire range. (Those deviations that are visible are largely due to our simplifying parameterization of the N-stage of our model as a simple cumulative Gaussian, and are not a failure of the LNP framework more generally.)

One very straightforward way of summarizing the quality of our model's predictions is to ask what percentage of each animal's choices we can guess correctly by looking at the model's output. For example, imagine that the model predicts that the monkey is more likely to choose red than green on the trial 653 given the events that led up to that trial. If our model is a good one, then we should guess that the monkey will in fact choose red. Using this simple rule, our LNP models correctly predict 80% of Monkey F's free choices and 79% of Monkey G's free choices—a surprisingly high rate given the stochastic nature of animal choice. (Again, these prediction rates are leave-one-out cross validated so that the data being predicted are not included in the set used to estimate the model parameters used to make the prediction.)

A slightly more sophisticated metric for summarizing the predictive performance of our model is the statistical likelihood of the actual behavioral data, given the predictions made by our model. We report this metric as a per trial, or average, likelihood. A detailed description of the computation of average-likelihood can be found in the Appendix (Equation A4), but the intuition behind this metric is straightforward. An average-likelihood near 1.0 indicates that the model made very strong predictions about animal choice on most trials *and* these predictions were largely correct (e.g., the model predicted that the animal was 99% likely to choose the red target on trial 653, and the animal did indeed choose red). An average-likelihood near 0.5 indicates that the model made only very weak predictions about animal choices (e.g., the animal was equally likely to choose red or green on trial 653). An average-likelihood near 0.0 indicates that the model made very strong predictions, but that these predictions were inaccurate (e.g., the model predicted that the animal was 99% likely to choose the *green* target on trial 653, when actually the animal chose red). Thus the advantage of average-likelihood over the more simple rate of correct prediction is that average-likelihood correctly penalizes models for overconfident erroneous predictions, but rewards them for making strong predictions that are also accurate. Again using leave-one-out cross validation, we found that our LNP model yielded an average-likelihood of 0.65 for Monkey F and 0.66 for Monkey G. Given the stochastic nature of the task, where events are never probability 1.0, these average-likelihood rates are very high. For all plausible behavioral models, even when data are artificially produced in simulation (as detailed in the next section) average-likelihood rates are below 0.70 between the synthetic choice data and the LNP model that actually constructed them.

Generative Sufficiency

We now turn to our second, and more critical, validation: testing the model's generative sufficiency. By this we refer to the model's ability (or inability) to independently generate behavior that is similar to the animals' in response to the changing reward contingencies of our dynamic matching task. Importantly, predictive sufficiency like that documented in Figure 6 is no assurance of generative success. Consider, for example, a television weatherman who simply predicts each day that there will be ''no change'' in the weather tomorrow. Because of temporal correlations in weather patterns, this algorithm

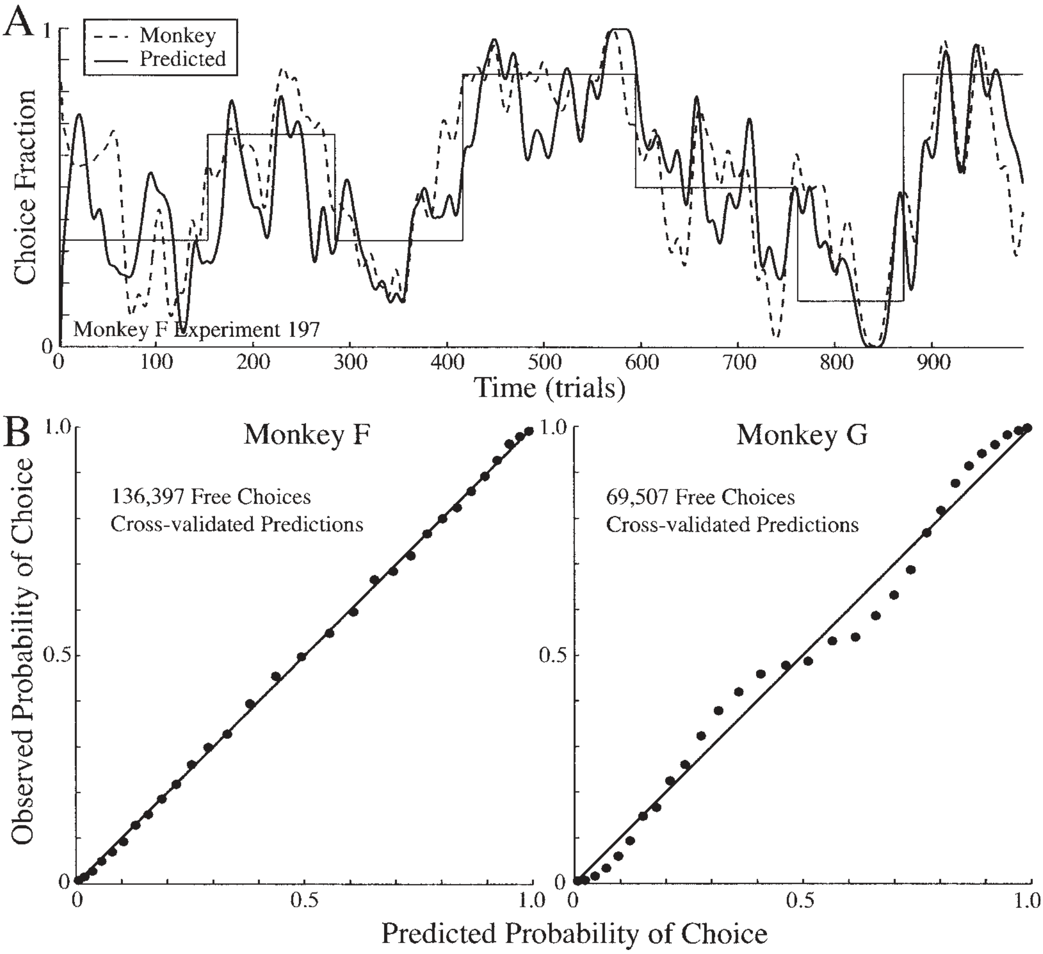


Fig. 6. Predictive performance of the model. (A) The monkey's local choice fraction for this experiment is replotted from Figure 2B now as a dashed curve. The solid line shows the choice fraction predicted by the LNP model diagramed in Figure 5, using the parameters given for Monkey F in Figure 4. Predicted choice fractions are smoothed with the same filter used for the actual choice fractions, shown in the inset of Figure 2B. For reference, the experimenter-manipulated fractional baiting probabilities also are replotted from Figure 2B as the same thin solid line. (B) Overall predictive performance of the model. Using the model recovered for each monkey, we computed a predicted probability of choice on each free-choice trial in our entire data set. Based on these predictions, trials were then sorted into 30 equally populated bins. In each bin, we compute the probability of choice actually observed as the fraction of red target choices in each bin. The 95% confidence intervals on these estimates are smaller than the plotted points.

can *predict* the weather correctly on an impressively high proportion of days. The algorithm incorporates no understanding of the mechanisms underlying weather production, however, and will fail miserably at *generating* realistic patterns of weather change-over time.

To test the generative sufficiency of our model, we simulated entire experimental sessions in which the model itself performs the foraging task, making choices and receiving feedback in the same manner as our monkeys. For each simulated experiment, we challenged our model with the identical sequence of baiting probabilities faced by the monkeys, and evaluated the model's performance just as we evaluated the monkeys' performance. In the course of these simulations, the model generates its own history of choices and rewards, which in turn provides the inputs that guide the model's subsequent choices—in contrast to the tests of predictive sufficiency above, the monkey is now completely ''out of the feedback loop.''

Figure 7A shows the results of simulated behavior for the same sequence of baiting probabilities shown in Figures 2B and 6A. We set the free parameters of the LNP model to the values recovered for Monkey F. We then allowed the model to interact with the foraging environment, choosing options and receiving feedback just as the animal would. For simplicity, we forced the simulated player to obey the COD rule rigorously; thus the LNP model was consulted to guide free choices only. The sequence of simulated rewards and choices generated by the model are plotted in the same manner as the animal data shown in Figure 2B, including smoothing with the same Gaussian kernel shown therein.

The qualitative similarity between the choice behavior of the monkey (Figure 2B) and the simulated choice behavior generated by the LNP model (Figure 7A) in this example experiment confirms the generative sufficiency of our model. In both cases, behavior adapts rapidly to block transitions and tracks local noise in the sequence of experienced rewards. In particular, the solid choice trace follows the dashed reward trace in 7A aggressively, as was evident in the monkey data shown in 2B. This close tracking of reward fraction by choice fraction was typical of all simulated experiments. The most substantial discrepancies between monkey and model behavior occur at the extremes of probability, which the monkey skirts cautiously but the model approaches boldly. We would better capture this aspect of the monkeys' behavior if the model included a modest bias toward exploring both options equally.

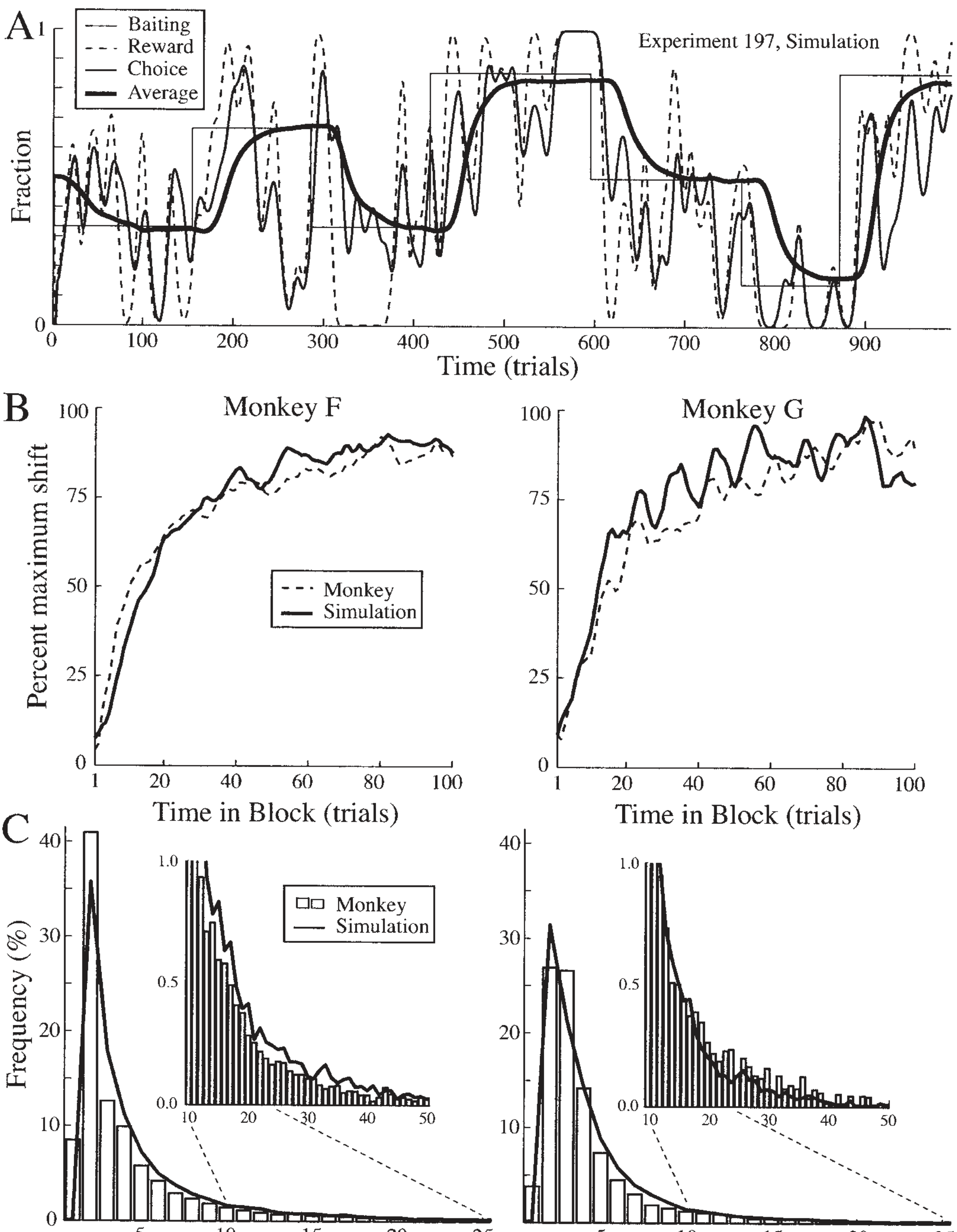
Though they are qualitatively similar, we expect the traces in Figures 2B and 7A to differ in their detailed structure. Even if our model were a perfect descriptor of the animal's strategy, the precise sequence of choices made in simulated and real behavior has to differ because the decision stage of the LNP model is probabilistic—we would expect the same degree of difference if we ran the animal on the same block sequence on 2 consecutive days. By simply repeating our simulation, we can also consider the average behavior of the model on this sequence of blocks. The double-thickness trace in Figure 7A shows the model's smoothed choice fraction averaged across one thousand such simulations in which we used different seeds for the random number generators that drive the P-stage of the model player and the probabilistic baiting of targets in the task.

Figures 7B and 7C depict tests of our model's generative sufficiency across the entire data set. Figure 7B focuses on the dynamics of behavioral adaptation at block boundaries. These plots show the time-course of choice fraction adaptation, computed exactly as for the behavioral data in Figure 2C. In the left panel, the solid line plots the adaptation time-course for synthetic data produced by our LNP model, using Monkey F's parameters. For comparison, the dashed line illustrates Monkey F's adaptation time-course taken from Figure 2C. The right panel shows the same comparison between synthetic data and real data for Monkey G. The correspondence for both animals is excellent, demonstrating that our LNP model is capable of independently generating realistic choice dynamics.

Figure 7C compares run-length histograms for synthetic and real behavioral data. In these panels, we plot the distribution of run lengths extracted from each monkey's real behavior, as we did for Monkey F in Figure 2D. For comparison, the thick solid line plots the distribution of run lengths that result from the model's simulated behavior. The agreement between real and simulated run lengths

→

Fig. 7. Generative performance of the model. (A) This panel follows the conventions established in Figure 2B, but now shows local reward and choice fractions for synthetic behavior generated when the model diagrammed in Figure 5 performed the foraging task in simulation. The additional double-thickness line shows the average choice fraction over many repetitions of this block sequence using different random number seeds, thereby averaging out the noise in individual runs. (B) Adaptation dynamics. The curves in this panel are calculated as per Figure 2C. The dashed lines show the time-course of choice fraction adaptation for each monkey after a block boundary transition. The thick black lines show the same time-course computed from synthetic choice data generated by the model we recovered for each monkey; simulations were run on the same block sequences experienced by each monkey during our experiments. (C) Run length histograms for each animal, following the conventions established in Figure 2D. The thick black lines show the distribution of run lengths produced in the same monkey-specific simulations used to generate the data in Figure 2B. The inset shows these distributions in finer detail for very long run lengths.



is excellent. The overall shape of the run-length distributions is slightly different for the 2 animals, but these differences are recapitulated in the shape of the distributions produced by the respective LNP model for each animal.

The most serious discrepancy between observed and simulated run lengths occurs at run lengths of one or two. We believe that this effect is attributable to our animals' imperfect adherence to the changeover-delay (COD) rule. By construction, the model adheres perfectly to the COD, never choosing a target only once—thus the frequency of runs of length one generated by the models is exactly zero. The animals, of course, do not obey the COD rule perfectly; they occasionally choose a color only once, resulting in a nonzero frequency of runs of length one. Similarly, for the animals, the effect of the COD may linger to influence the frequency of run lengths of two. For a run length of two, the model is completely guided by our LNP model, and never misapplies the COD rule on these trials. For the animals, we have no guarantee that the distinction is this crisp. For example, Monkey G, that obeys the COD rule more strictly at run lengths of one, may occasionally lose track of how many times it has chosen the current target, allowing the COD rule to influence its decision at a run length of two.

For these reasons, we believe that longer run lengths, those far from the influence of the COD, are the best test of our model. Correspondence is quite good at run lengths of three and higher. To show the strength of this correspondence, the inset depicts a magnified view of the distributions for very long run lengths. These events are very rare, each making up a fraction of a percentage of the entire data set. Even at these extremes, however, the monkeys' distributions follow the shape of those generated by the models. The magnitude of the noise in the histograms even appears to be similar. (The overall height of the histogram at long run lengths is influenced by the frequency of run lengths of one or two because frequency histograms are normalized to have equal area—it is the shape and not the absolute levels that are most important to compare.) The correspondence in distribution shape, even at this level of detail, strongly supports our earlier suggestion that run lengths produced by the monkeys result from a stochastic choice mechanism. The model is explicitly probabilistic in the P-stage and predicts that we should occasionally observe 50 consecutive choices to a single color (rightmost histogram bin in the inset); that these events actually appear in the data set at a frequency similar to that predicted by the model is remarkable, and strongly supports the long-standing claim that animal choice is probabilistic in the context of matching tasks.

Just as we summarized our model's predictive performance as the percentage of each animal's choices correctly predicted, we can summarize generative performance using these run-length histograms. To do this we compute the *overlap* between the run-length histogram observed in each animal and the run-length histogram generated by the recovered LNP model for that animal. The overlap of two distributions is defined as the percentage of the area under one histogram that is also under the other. If the model perfectly described the animal's choice behavior, the two histograms would be identical, and thus have a 100% overlap. As the statistics of run-length predictions shift away from those observed in the animal, overlap decreases. For Monkey F, the reconstructed model yields an 87% overlap between observed and generated run length histograms, and 91% overlap for Monkey G.

The N-Stage: Difference or Division?

A major difference between the LNP model shown in Figure 5 and the model in Sugrue et al., (2004) is the form of the N-stage. In this previous work, the N-stage was based on simple division, invoked by analogy to the classical matching law:

$$pc = \frac{v_r}{v_r + v_g}, \tag{16}$$

where v_r and v_g are the values of the red and green options, which we operationalized above as $L[r_r(t)]$ and $L[r_g(t)]$, the output of our L-stage operating on the binary reward history vector for each color. For obvious reasons, the expression on the right side of this equation was termed *fractional value*, which served as the scalar value metric in the previous study.[2] In contrast, the N-stage used here is a nonlinear function of our scalar value metric, *differential value*:

$$pc = s(v_r - v_g), \tag{17}$$

[2] Sugrue et al. (2004) actually used the term *local fractional income*. In that study, the L-stage linear filter was assumed to be a leaky integrator (i.e. a pure exponential filter), and this term emphasized that the inputs to the nonlinearity were summed rewards, or income. Here we assume no particular form for the L-stage filter (indeed, in the next section we even relax the assumption that it operates only on rewards), so we are forced to abandon the specific term *income* in favor of the more general concept of *value*.

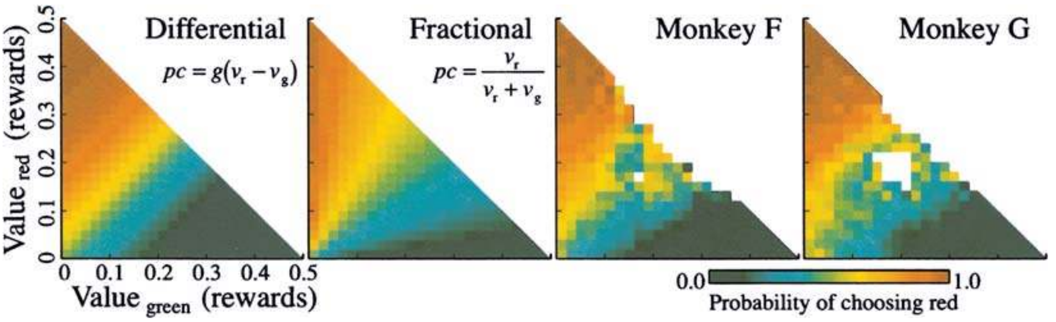


Fig. 8. Comparison of candidate N-stage nonlinearities. Each panel shows a probability of choice surface as a function of the linear filter outputs for both targets. Thus the pixel in the upper left corner of each panel shows the probability of choosing the red target when the filtered reward history on the red target produces a value of 0.5 rewards and the filtered reward history on the green target produces a value of 0.0 rewards. The leftmost panel shows the probability of choice surface produced by the model diagrammed in Figure 5, where a sigmoidal nonlinearity operated on the difference of filter outputs (i.e., differential value). The second panel shows the probability of choice surface produced by a model that computes probability of choice by expressing the output of one filter as a fraction of the summed outputs of both (i.e., fractional value). The right two panels show the observed probability of choice surface that was actually observed for each of the 2 monkeys in our study. In these two panels, the filter output values were computed using the filter recovered for that particular animal (Figure 4A). Bins containing fewer than 10 free-choice trials are shown in white.

where $s(x)$ is the sigmoidal function that smoothly maps differential value to 0 or 1.

Which of these two nonlinearities better characterizes the choice behavior of our animals? To make this comparison, we consider the surface defined by the two-dimensional function:

$$pc = f(v_r, v_g), \tag{18}$$

where f is either of the two candidate N-stage nonlinearities. Intuitively, we simply plot the probability of choosing red predicted by each of the two nonlinearities for every combination of v_r and v_g. The leftmost two panels of Figure 8 show these predicted probability surfaces for the differential value nonlinearity and fractional value nonlinearity, respectively. The contours of constant probability of choice, represented as constant colors in the figure, differ strikingly for the two models. The contours generated by the differential value nonlinearity are parallel, whereas those generated by the fractional nonlinearity radiate from the origin like the spokes of a wheel.

We compare these alternative predictions with direct measurements of the probability of choice for each of our monkeys, shown in the two rightmost panels of Figure 8. These *two-dimensional* surfaces are generated in a similar manner as the *one-dimensional* curves in Figure 4B. For each free-choice trial, we simply compute v_r and v_g, using the L-stage filters for each animal, and then bin the animals' actual choices based on these values.[3] Within each of these bins, the fraction of free choices made to the red target provides us with the maximum likelihood estimate for the probability of choice under these conditions, which we show color coded. White gaps represent bins with insufficient data (less than 10 free choices to either color).

The plots reveal that the nonlinearity based on differential value employed in the current study describes the choice behavior of both animals better than the fractional value nonlinearity used in our earlier paper. Contours of constant probability of choice are nearly parallel for both animals, strongly suggesting that the decision process depends on the difference of the L-stage filter outputs rather than their ratio. This result is surprising because the fractional-value-based model explicitly is a local implementation of the matching law. In fact, the best mechanistic

[3] To generate the two rightmost panels of Figure 8 we used the filters recovered for the L-stage of our model to compute the value of each color. Some might worry that this would bias us in favor of the N-stage nonlinearity that we also recovered, and that any apparent preference for differential versus fractional value is induced by our choice of filter. To ensure that this was not the case, we repeated the analysis with arbitrary filters. For both a single exponential and a simple box filter the clear preference for differential value remained.

description of our animals' matching behavior (differential value) bears little resemblance to Herrnstein's original formulation of the matching law.

We might consider an even more elaborate model, employing both operations sequentially. In such a transformation (which we could term *differential fractional value*) value signals first would be normalized by division (as fractional value), and subsequently compared using subtraction (as differential value). Unfortunately, no explanatory power can be gained by adding this complication. Algebraically,

$$\frac{v_r}{v_r + v_g} - \frac{v_g}{v_r + v_g} = 2\left(\frac{v_r}{v_r + v_g}\right) - 1, \quad (19)$$

where the left side of the equation is differential fractional value and the right side is a trivial linear transformation of fractional value. This identity follows from the fact that there are only two options that the animal is choosing between in our task: red and green. Thus the analysis above, which strongly favored differential value over fractional value, will similarly favor a pure differential model over models that compute differential fractional value as an intermediary.

Choice of Inputs

What input to the LNP model is most appropriate? Recall that we elected to use a binary string of ones and zeros coding the delivery or absence of a reward on each trial. In essence, this selection of inputs implies that only rewards exert a significant influence on the monkeys' subsequent choices. But this assumption is, as yet, untested. It might be, for example, that the history of past choices also conveys useful information about the animals' future actions.

One naive approach to this question is to construct an explicitly *choice-based*, rather than *reward-based* model. To construct this model, we supply choice history rather than reward history as the raw binary input and repeat the estimation procedures described above. As a result, we recover a different L-stage filter (Figure 9A) and a different N-stage decision criterion (Figure 9B)—in effect, an entirely different model of choice behavior. As shown in Figure 9C, this choice-based model is actually a surprisingly good predictor of animal choice. This simple model correctly predicts 77% of Monkey F's free choices (as compared to 80% for the reward-based model) and has an average-likelihood of the actual data given the model predictions of 0.60 (as compared to 0.65). The counterintuitive success of the choice-based model arises from the fact that a monkey that is matching well will continue choosing between the two targets in roughly the same ratio within a block of trials. Thus information about the recent ratio of choices within a block permits predictions about the future ratio of choices even without direct access to reward information.

Although the choice-based model performs reasonably well in predictive testing, it is completely inadequate in generating realistic behavior, as is shown in Figure 9D. By design, this model ignores all reward feedback from the outside world and bases its future actions on its own previous behavior. Thus it wanders randomly, choosing one target more or less frequently, and is completely blind to changes at block boundaries. The double thickness line in Figure 9D shows the average behavior of this model over many simulations with different random number seeds. Plainly, this average choice behavior is completely uncorrelated with shifts in baiting fraction (compare with Figure 7A). (Correlation coefficients between smoothed choice fraction and baiting fraction for both real data and synthetic data produced by the reward-based model are approximately 0.7 or 0.8, whereas the corresponding correlation coefficient computed for this choice-based model is less than 0.02 over the entire data set.) The small bias seen in the average trace is a product of a similar bias present in the recovered nonlinearity in Figure 9B.

Although choice history alone cannot be used as a basis for building a valid model of behavior, perhaps we can combine information from both reward and choice history to construct a better model than either alone. In our original model, the input vector for red contains zeros for all trials where red was not rewarded, regardless of whether red was *chosen and not rewarded*, or red was simply *not chosen.* Intuitively, however, these events might have very different effects on subsequent choices: An unrewarded choice might in fact motivate the monkey to switch options, actively driving the value of the unrewarded option down rather than only reducing value through

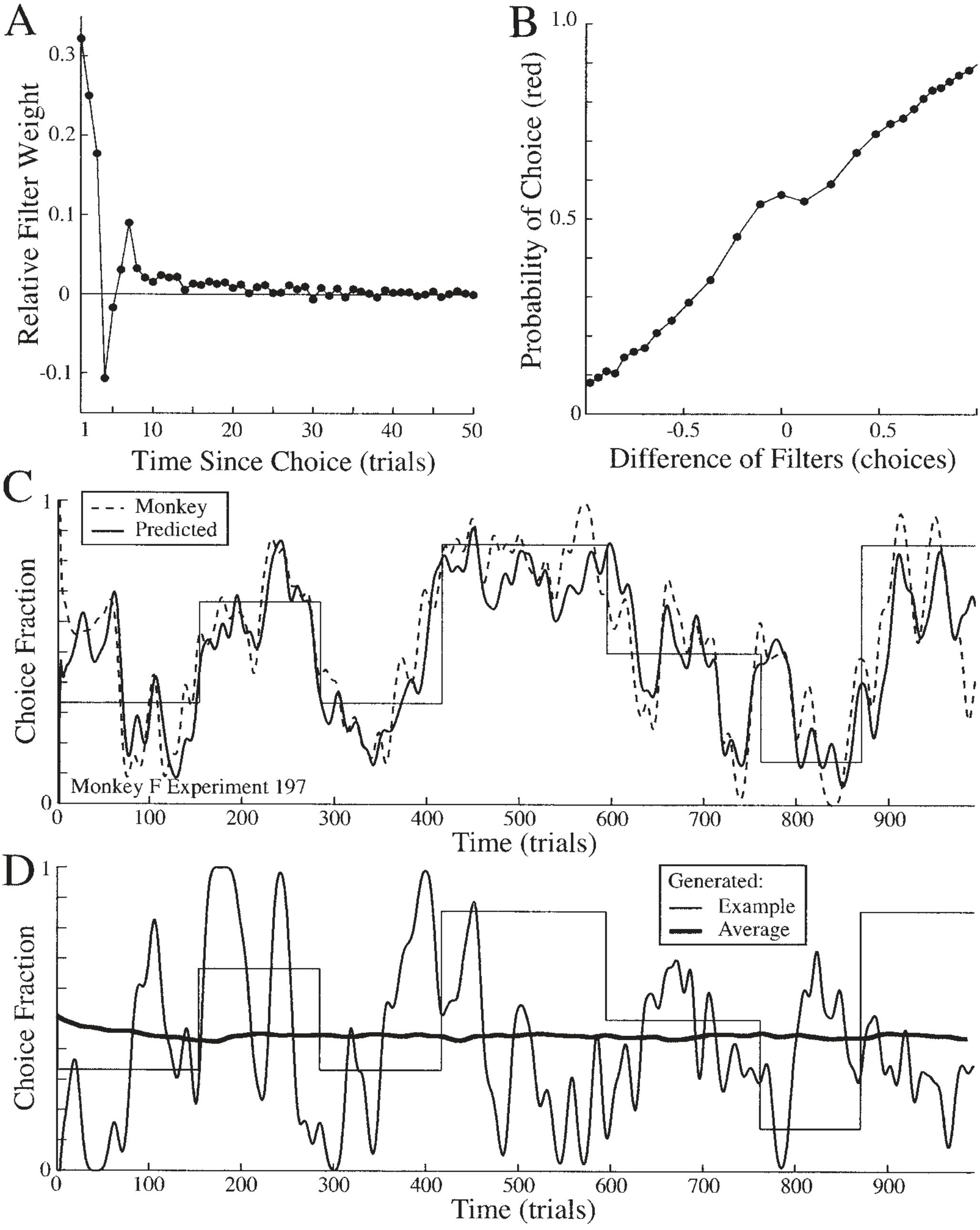


Fig. 9. A choice-based model—an LNP model recovered for Monkey F that relates past choices, rather than past rewards, to future choice. (A) L-stage filter, analogous to Figure 4A. (B) N-stage decision criterion, analogous to Figure 4B. (C) Predictive performance, analogous to Figure 6A. (D) Generative failure, analogous to Figure 7A.

passive decay. If this intuition is correct, we might predict choices better by constructing an input vector where rewarded choices to a color are coded as positive entries, but unrewarded choices to that color are coded as negative entries.

We tested this type of reward-choice ''multiplexing'' by examining the predictive and generative performance of models based on input strings coded with ones for rewarded choices, δs for unrewarded choices, and zeros if the target was not chosen. Positive δs would

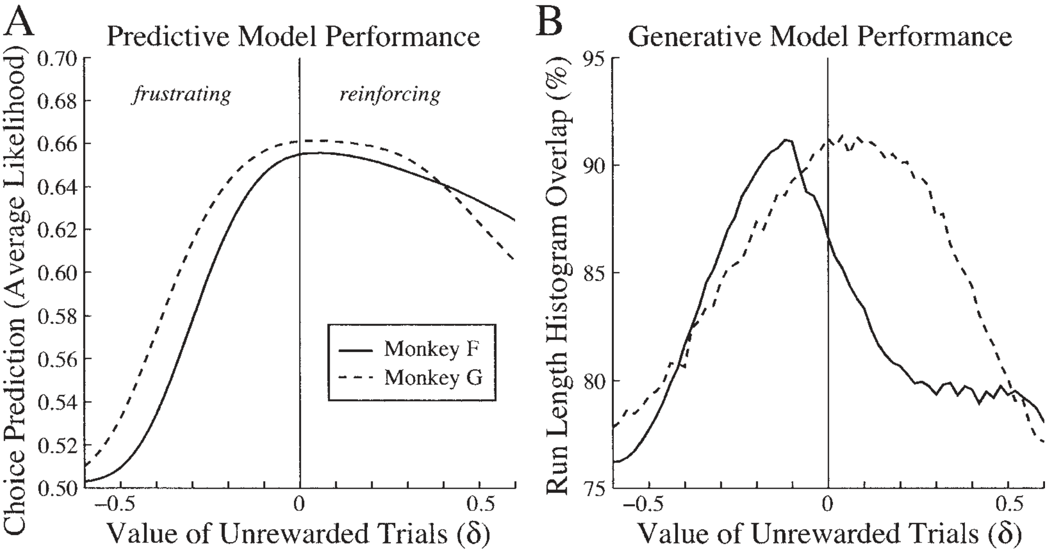


Fig. 10. Effect of unrewarded trials on model performance. Rather than coding reward histories with ones for rewarded trials and zeros for all unrewarded trials, we code reward histories for each color with ones when that color was chosen and rewarded, δs when it was chosen and unrewarded, and zeros when it was not chosen. Positive δs indicate that unrewarded choices are reinforcing (having the same sign as rewards), whereas negative δs are frustrating (having the opposite sign). For each value of δ, we reestimated the LNP model for each monkey. (A) Effect of δ on the prediction of free choices for each monkey. (B) Effect of δ on the overlap between observed and generated run-length histograms for each animal.

imply that unrewarded trials actually reinforce the chosen option (as does a reward), whereas negative δs mean that unrewarded trials decrease the apparent value of the chosen target, actively motivating the animal to switch—as suggested by the intuition developed above. Figure 10A shows the effect of varying δ on the predictive success of the model. For each value of δ plotted, we reestimated the best LNP model for each monkey and report the average-likelihood, $\delta = 0$ being the original value-based model shown in Figure 4. (Results are similar if we plot the percentage of free choices correctly predicted.) Contrary to the intuition developed above, negative values of δ offer no improvement and actually impair predictive success—dropping prediction rates to chance levels by $\delta = -0.5$. This suggests that unrewarded choices do not provide an active incentive to switch targets. In the other direction, positive values of delta offer only infinitesimal gains in predictive performance before rolling off into more modest but notable decreases in predictive performance. In the limit where $\delta = 1$ (all choices to a color are equally reinforcing regardless of whether they are rewarded), this reduces to the pure choice-based model illustrated in Figure 9.

Generative testing does not strengthen the case for the usefulness of these multiplexed models. Figure 10B plots the overlap between the run-length histograms for each monkey and the run-length histograms for synthetic data generated using each particular value of δ. Both curves show significant falloffs as values of δ grow either more positive or negative. In 1 monkey, an approximately 4% increase in overlap can be achieved by using a value of δ near −0.1, but absent a corresponding result in the other monkey and at the cost of as nearly as large a decrease in average-likelihood, this modification to our original model is unwarranted. Thus the data provide no compelling reason to replace the simple reward-based model with a δ-based model.

Still more elegant methods exist for combining information from reward and choice in the hopes of constructing a better model of behavior (see, e.g., Lau & Glimcher, 2005). In this vein, we have explored the predictive and generative success of a model that includes both reward history and choice history as separate inputs. We considered LNP models that allow distinct L-stage filters for rewards and choices, the outputs of which are in turn linearly combined before reaching the N-stage. To accurately reconstruct such a model from data, the appropriate kernels for filtering reward and choice information must be estimated *simultaneously* so that information already effectively captured by one is not reiterated by the other. This can be accomplished by generalizations of our above Wiener kernel approach that allow for the simultaneous extraction of correlations with multiple input streams.

This ''joint'' reward/choice approach produces LNP models that perform marginally better than our reward-based model in predicting future choices *given* the animal's recent actions (average likelihoods of 0.66 for both monkeys), but the generative performance of these models is slightly worse (80% and 86% overlap in run length histograms for Monkeys F and G, respectively). Thus the more complex models offer little or no advantage in understanding behavior in our task. Lau and Glimcher (2005), using a similar joint modeling approach, obtain more substantial improvements in predictive performance with the incorporation of choice information. A likely reason for this discrepancy is that Lau and Glimcher did not employ a changeover delay to punish alternation between the two targets. Any residual tendency toward alternation would introduce predictable regularities into sequences of choices independently of reward history, which a joint model could easily recover. For our data set, extraction of additional information from the history of choices and rewards will require more general models that allow nonlinear interactions between reward and choice.

Optimality of Model Parameters

Having validated our LNP model through generative testing, we now leverage that same simulation environment to examine the effect of the various free parameters on foraging performance. In turn, we use this information to evaluate how closely the monkeys approximate an optimal behavioral strategy within the subclass of LNP models diagramed in Figure 5.

We explored a wide range of values for each of the parameters in our LNP model but focused on three parameters in particular: τ_1 and τ_2, the short and long timescale components of the L-stage linear filter, and σ, the width of the N-stage sigmoidal nonlinearity. We enforced the constraint that $\tau_2 \geq \tau_1$ to preserve their interpretation as the short and long timescale components. The effects of varying the mixing coefficient for the short and long L-stage components, *a*, and the residual N-stage bias, μ, are discussed briefly below. For each combination of parameters tested, our LNP model performed 10 repetitions of the entire block sequence each animal experienced during its experimental history, resulting in a total of over 2.5 million simulated trials per parameter combination.

We quantify task performance in terms of *harvesting efficiency*, the average number of rewards earned per trial. We report this metric as a percentage of the combined baiting probability on both targets. In these units, a harvesting efficiency of 100% is, in fact, impossible to achieve because it would require that every reward be gathered immediately when placed, even if two targets were baited simultaneously. At the other bound, a harvesting efficiency of 50% would be achieved by choosing only one of the two colors for the course of each day's experiment—because, averaged across all experiments, rewards are equally likely to appear on either red or green.

The three panels of Figure 11A show cross sections through the parameter space of the simulations, each showing the effect on harvesting efficiency of varying a single parameter. The leftmost panel shows the effect of varying τ_1. For each value of τ_1, we optimized the choice of τ_2 (with $\tau_2 \geq \tau_1$) and σ to maximize harvesting efficiency; thus each data point in this plot reflects the best harvesting efficiency possible for any model with the specified τ_1. Arrows indicate the values of τ_1 displayed by Monkey F and Monkey G. Similarly, the center panel shows the effect of fixing the value of τ_2 at various levels while optimizing τ_1 and σ to maximize performance. Finally, the rightmost panel of 11A shows the effect of manipulating σ, the width of the sigmoidal N-stage nonlinearity, while optimizing τ_1 and τ_2. For all simulations in 11A and B, *a* was fixed at 33% and μ was fixed at 0.0 rewards.

Figure 11A demonstrates that harvesting efficiency is substantially more sensitive to the precise values of τ_2 and σ than to τ_1. The sensitivity to τ_2 is easy to understand. In

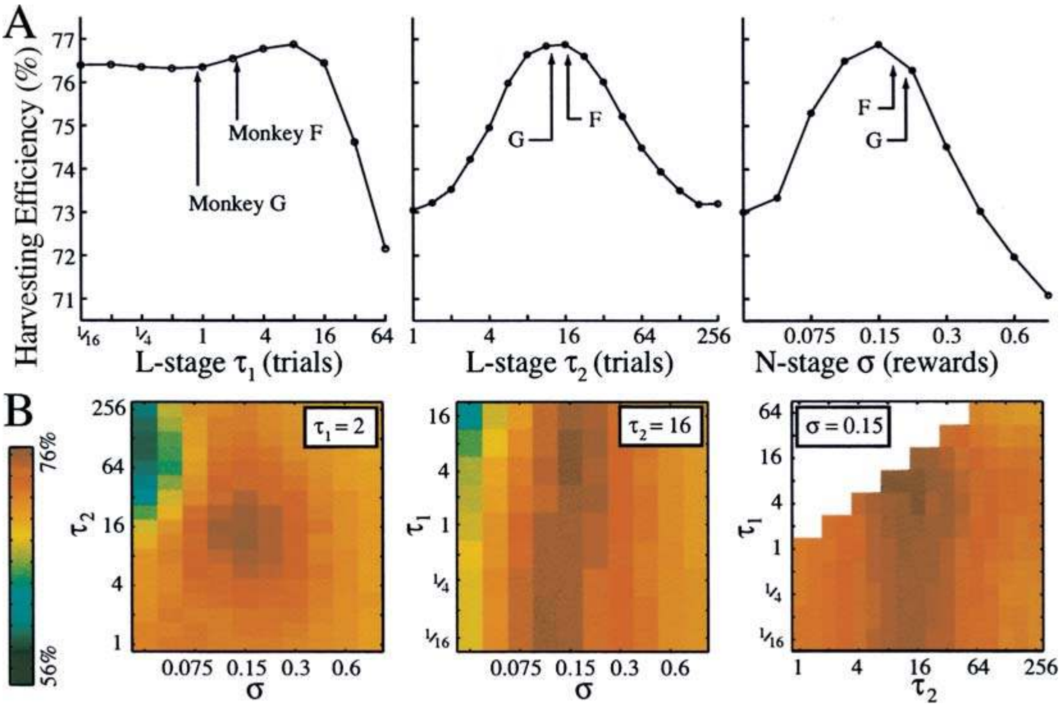


Fig. 11. (A) Effect of model parameters on reward harvesting. From left to right, the three panels show the effect of varying τ_1, τ_2, and σ on reward harvesting efficiency. Harvesting efficiency is defined to be the average rewards earned per trial, normalized by the total baiting probability on both targets. Each point shows the harvesting efficiency of the best-performing model using the specified parameter value. Thus the leftmost point in the leftmost panel is the performance of the best model using a τ_1 of 1/16 of a trial, whereas τ_2 and σ are free to assume any value. (Note: Very small values of τ_1, e.g., 1/16, approximate the case where only the first trial is given nonzero weight.) Each point represents the average harvesting efficiency of this best-performing model over 10 repetitions of our entire data set, using different random-number seeds for each repetition. Arrows indicate the parameter values recovered in Figure 4 for each monkey. (B) Cross sections through the volume showing harvesting efficiency as a function of all three parameters. From left to right, the three panels show a plane with constant τ_1, τ_2, and σ, respectively. The value of the fixed parameter is shown in the inset. Harvesting efficiency for all combinations of the other two parameters is shown color coded from blue (poor harvesting efficiency) to red (high harvesting efficiency).

essence, the L-stage of the model estimates reward rate by integrating reward history over a timescale governed by τ_2. This timescale of integration must be long enough to smooth out meaningless temporal variation induced by infrequent, stochastic reward events. It must be short enough, however, to permit rapid changes in choice behavior in response to real changes in the state of the world (i.e., block transitions). In our task, therefore, the overall reward probability and the frequency of block transitions conspire to define a narrow range of integration time constants—centered around 15 trials—that convey useful information about actual reward rates. Model performance declines precipitously if τ_2 differs from the optimal value by more than a factor of two in either direction, and this decline cannot be rescued by manipulating either τ_1 or σ. As indicated by the arrows, the animals' τ_2 parameter values are tightly grouped near this maximum in harvesting efficiency.

The same effect is evident for τ_1 (Figure 11A, left panel), where optimal performance also results from values near 15 trials. Values of τ_1 greater than 15 trials catastrophically reduce harvesting efficiency, because this manipulation also displaces τ_2 from its ideal value under our constraint that $\tau_2 \geq \tau_1$. Interestingly, there is very little penalty—1% or less—for choosing values of τ_1 much smaller than optimal. In fact, because the optimal tuning of time constants is $\tau_1 \cong \tau_2 \cong 15$, a single exponential temporal filter with this time constant would perform just as well as any of the double exponential filters we consider. For reasons that are not clear, neither of our animals use this simpler weighting function; both seem to be disproportionately influenced by the outcome of the most recent trial or two (τ_1). The animals pay little penalty for this strategy, but neither do they achieve any gains, a point that we consider further in the Discussion.

In addition to τ_2, performance depends critically on σ, the width of the decision criterion at the N-stage of the model (Figure 11A, right panel). Substantial penalties result from values of σ that depart by a factor of two or more from the optimal value of 0.15 rewards. The cost of a very shallow decision criterion (large σ) is obvious: the model fails to respond appropriately to substantial evidence favoring one or the other choice. In the limit, when σ is very large, the model chooses each option with near 50% probability regardless of reward history. The disadvantage to using very steep decision criteria (small σ) is more subtle, and depends upon our use of VI schedules, a point to which we return in the Discussion. Again, both animals seem to have tuned their choice of σ to optimize reward harvesting performance.

Figure 11B explores potential interactions between model parameters that might be concealed in Figure 11A (recall that each data point in 11A reflects optimization of the other two parameters). To generate the three panels in Figure 11B, we fixed one parameter at a typical value and calculated harvesting efficiency for all combinations of the other two parameters. The data reveal no substantial interactions between parameters, confirming the accuracy of the optimality tuning portrayed in 11A for τ_1, τ_2, and σ. For example, in the leftmost panel of Figure 11B, with τ_1 fixed at two trials (close to the monkeys' τ_1 values), we see a clear maximum at the values of τ_2 and σ suggested in 11A, with harvesting efficiency falling roughly symmetrically in all directions immediately around the peak near the center of the graph. Similarly, the center and right panels of Figure 11B confirm that for any particular choice of σ and τ_2, respectively, the choice of τ_1 matters relatively little, as was suggested by the left panel of Figure 11A. This lack of influence of τ_1 is revealed by the relatively constant vertical bands of color in the panels in question. Such comparisons indicate that the optimization of model parameters is nearly (though not perfectly) separable, thereby validating the simpler single-dimensional approach to parameter exploration shown in Figure 11A.

The effects of varying the residual N-stage bias, μ, and the mixing coefficient for the short and long L-stage components, *a*, are relatively straightforward. The bias term μ introduces a static preference for the model to choose one or the other option regardless of reward input. Any value of μ that deviates from zero results in a decrease in harvesting efficiency; both monkeys exhibited bias values very close to optimal (0.02 rewards for Monkey F and 0.01 rewards for Monkey G). Altering the mixing coefficient, *a*, alters the relative impact of tuning for τ_1 and τ_2 in an intuitively predictable manner. If the value of *a* is very small, all the filter weight is put on τ_2 and optimizing τ_1 becomes completely irrelevant. If value of *a* is larger, the pressure to optimize τ_1 becomes greater, although the need to tune τ_2 persists until *a* approaches 100%. Thus varying *a* determines *which* of the other parameters must be tuned to obtain good performance, but itself has little effect on harvesting efficiency.

DISCUSSION

As in Herrnstein's seminal papers, the aim here has been to elucidate the relation between past reward and future choice. To this end, we have proposed a mechanistic model of choice that describes the behavior of primates in a dynamic foraging environment (Figure 5). Working within the broad framework of LNP models, we reconstructed this behavioral model directly from the data. First, using a generalization of Wiener kernel analysis, we recovered an L-stage linear filter that captures the relative influence of recently experienced rewards. For both monkeys, this impact of past rewards is strikingly well described by a double-exponential decay (Figure 4A). These recovered linear filters allowed us to compute an intermediate scalar value metric, which we were able to relate directly to probability of choice through our N-stage nonlinear decision criterion. This relation, recovered by maximum likelihood estimation, reveals how the comparison of option values impacts an animal's distribution of choices. For both animals, this relation is well described by a sigmoidal function (Figure 4B) operating on the difference between the two option values (Figure 8). In simulations, the complete LNP model was effective both in predicting future animal choice from recent behavior (Figure 6) and in generating realistic synthetic choice behavior when acting independently (Figure 7).

The parameters of the foraging strategy adopted by our animals appear nearly optimal among the subclass of LNP models shown in Figure 5. The complete model has five parameters: the short and long terms of the double exponential weighting function (τ_1 and τ_2), a mixing coefficient describing the balance between the two terms of the exponential (a), and the mean and standard deviation of the cumulative normal distribution underlying the sigmoidal nonlinearity (μ and σ). The tuning of parameters that have a strong influence on harvesting efficiency—τ_2, σ, and μ—is very similar between the animals and is extremely close to the values that achieve maximum rewards earned per trial. Meanwhile, the tuning of parameters that have less influence on harvesting efficiency—τ_1 and a—is less stereotyped but at little cost to achieved performance. This strongly suggests that the behavior of the animals adapted during extensive training to become optimized for the statistical structure of reward availability in our task.

Two groups previously have reported relatively rapid shifts in matching behavior in situations in which animals have been exposed to frequent changes in reward contingencies. During daily experimental sessions, Gallistel and colleagues exposed rats to a single signaled (Mark & Gallistel, 1994) or unsignaled (Gallistel et al., 2001) change in reward schedules. Davison and Baum (Baum & Davison, 2004; Davison & Baum, 2000) exposed pigeons to seven changes per experimental session; however, the time of occurrence of each change was signaled by a preceding blackout period. Our animals experienced approximately 12 unsignaled changes in the relative baiting probabilities each day, marking an unprecedented level of uncertainty in reward schedules. The animals' behavior adapted rapidly to new reward contingencies (Figure 2C) in all three of these studies, a result that contrasts strikingly with the large majority of prior studies in which animals typically required many sessions to stabilize on a new schedule. It appears, therefore, that behavioral dynamics can be influenced strongly by the statistical structure of the reward schedules: rapid changes in reward contingencies generate rapid changes in choice allocation. Ultimately, animals seem to adopt a behavioral strategy that exploits the statistical regularities of a given environment (Killeen, 1981, 1994; see also Lau & Glimcher, 2005).

Heavy-Tailed Weighting of Rewards

The most puzzling aspect of the behavioral strategy revealed by our analysis is the stereotyped shape of the value weighting functions shown in Figure 4A. The shape of this linear filter is extremely well described as a mixture of two exponentials, yet the use of a double exponential value weighting filter cannot be motivated from the structure of our task. Optimal reward harvesting could have been obtained using a *single* exponential kernel with a time constant of approximately 15 trials. Although the animals tune the long timescale component (τ_2) of their filter very near to this optimal value, both animals also display a pronounced additional component of their linear filter at a much shorter time scale (τ_1). As discussed above, this ''extra'' weighting of very recent events costs the animal little in terms of harvesting efficiency, but this does not explain its presence. We believe the answer to this riddle is related to temporal discounting, a topic of intense interest in the field of behavioral economics.

We know from behavioral economics that when humans evaluate future rewards, more distant rewards are depreciated, or *discounted,* relative to less distant rewards (reviewed in Frederick, Loewenstein, & O'Donoghue, 2002, and Green & Myerson, 2004). Thus a reward of the same objective magnitude that is delayed in time is valued less than a comparable, but more immediate, reward. This sort of prospective discounting of rewards is rational under the assumption that increased temporal delay is associated with increasing uncertainty about whether the reward will actually occur. Normative economic models based on this assumption predict a constant rate of discounting for future times, leading to an exponential weighting profile for the value of future rewards (Strotz, 1956). However, experiments in behavioral economics suggest that humans and animals alike seem to employ *hyperbolic*, or superexponential, discounting (e.g., Green, Myerson, & McFadden, 1997; Kirby & Maraković, 1995; Loewenstein, 1987; Mazur 1987; for a review, see Green & Myerson, 2004). That is to say, experimentally measured value weighting functions fall off steeply at first, but then continue with a long tail. (The use of the term ''hyperbolic'' is originally due to Herrnstein, who was the first to use a hyperbolic form to describe the shape of prospective discounting functions in pigeons; see Chung & Herrnstein, 1967). Whether heavy-tailed discounting functions are truly hyperbolic, in the mathematical sense, or might be better described as a mixture of two single exponentials with different time constants, has not been convincingly demonstrated. For example, McClure, Laibson, Loewenstein, and Cohen (2004) employ double exponential discount functions to explain human behavioral and neural data.

Our situation is certainly different, but the analogy is striking. We are not investigating the discounting of *future* rewards, but we are measuring the discounting of *past* rewards. The linear filters we recover as the first stage of our LNP model are in fact a direct measure of the animals' value weighting for rewards at different time points in the past. Moreover, the motivation for discounting at all in our experimental setting stems from uncertainty in the world, just as it is for prospective discounting. In our case, it is not that receipt of some future promised reward may be thwarted by changing contingencies, but that information received about past rewards may be rendered irrelevant by changes in baiting probabilities at the block boundaries. Similarly, though a single exponential discounting function would be optimal, both of our animals show clear heavy-tailed, or hyperbolic, value weighting functions. (The value weighting functions in Figure 4A were significantly better fit by a double exponential form than they were by hyperbolic forms with a similar or fewer number of free parameters. Thus our use of the term hyperbolic is only meant to be qualitative, consistent with its usage in behavioral economics.)

These parallel findings suggest the interesting possibility that the representation of rewards in the brain may have an intrinsically hyperbolic character across time—that both prospective and retrospective rewards are typically discounted with heavy-tailed weighting functions. McClure et al. (2004) have suggested that the double exponential nature of prospective value weighting functions may reflect the existence of two separate evaluative processes within the brain operating at different time scales. Were these two subsystems operating to encode the value of retrospective as well as prospective rewards, the overall shape of the linear filters recovered for our L-stage might reflect a biological constraint. In this view, through learning our animals have tuned the overall timescale of this characteristic weighting function to exploit the statistics of reward delivery in our task, and the ''extra'' unaccounted for weighting of very recent times is a signature of the underlying neural mechanisms of valuation.

Importance of the Sigmoid N-Stage

In most discrimination or comparison tasks, it is advantageous to use as steep a decision criterion as possible. For example, to have the greatest chance of correctly identifying which of two light sources has greater luminance, one should always choose the option with the greater measured value even if the available measures come from very noisy sensors and the difference in the measured values is miniscule. For our task, however, performance is maximized not by the steepest possible decision criterion, but by one with a gradual transition. This graded transition between preferring the more and less valuable target requires the use of a sigmoidal decision function rather than a hard, binary decision criterion.

The disadvantage to using very steep decision criteria in our task hinges on our use of VI schedules. Because VI schedules leave uninspected options baited until they are chosen, it is in the player's best interest to occasionally visit the less valuable option to collect rewards that have accumulated there over time (Houston & McNamara, 1981). It is for this very reason that matching behavior is so nearly optimal for VI schedules: A player should not *exclusively* choose the option with the higher baiting probability, but instead distribute choices between both options, *favoring* the target with the higher baiting probability. This need for continuous titration of choice fraction as a function of baiting probability is the reason why a smoothly transitioning decision criterion is preferable.

It is essential to distinguish this situation from others in which smoothly transitioning choice functions are observed. For example, in sensory psychophysics it is common to see smooth transitions in choosing one option over the other. These ''psychometric'' functions also have a sigmoidal shape, but do not necessarily indicate that a smooth decision criterion is being used internally. For an internal decision criterion that is a perfect digital step function, noise at the level of the input sensors and the finite limitations of experimental control can still conspire to give the appearance of a smooth transition in the resulting behavior (Green & Swets, 1966). However, were one to construct an artificial decision maker with finite sensor accuracy, and parametrically vary the steepness of the internal decision criterion, one would see that any decrease in steepness results in poorer performance. We have found just the opposite. For an artificial player in our task, it is *advantageous* to use a soft or blurry internal decision criterion. Moreover, our animals have optimized the width of that decision criterion to maximize rewards—a finding that argues very strongly for the presence of an explicit smoothly varying internal decision function.

The Importance of Generative Modeling

We validated our model by testing both its predictive and generative sufficiency. First, in testing predictive sufficiency, we quantified our model's ability to predict the next choice the monkey would make, given the animal's recent reward history. Second, we explored our model's ability to generate behavior on its own in a simulated foraging environment and compared the resulting behavior to that of our animals. Given the obvious success of our model in cross-validated tests of prediction, is generative testing actually important? In fact, generative testing is absolutely critical because adequate predictive performance (compare Figure 9C to Figure 6A) can be coupled with catastrophic generative failure (compare Figure 9D to Figure 7A). Without the perspective of generative testing, the choice-based model that gave rise to the predictive results in Figure 6A might be considered more plausible than it actually is.

The key to why the results of predictive and generative testing can be so different is feedback. In the simulations used for generative testing, the choices that the model makes influence its own subsequent view of the foraging environment. In other words, the model's output feeds back into its input, a process that reflects its interaction with the world. In predictive testing, there is no such feedback; the model is only tested against conditions that were generated by the animal, not conditions that the model itself had an active role in shaping. No passive test of goodness of fit can reveal how a choice model will react under the active feedback scenario of generative testing. Thus generative ability emerges as a particularly incisive test for models of choice behavior, especially in contexts characterized by extensive feedback, where individual choices have consequences that inform future choices. Indeed, we propose that generative testing is essential for critical evaluation of any model of behavioral choice.

It is worth noting that the analytic techniques employed in this paper—Wiener kernel analysis and maximum likelihood estimation—are inherently biased toward finding better predictive, not generative, models. These techniques make no attempt to understand foraging behavior as a feedback system, but instead focus on relating input to output. More effective generative models might emerge in the future from approaches such as learned Hidden Markov Models, which can explicitly incorporate the feedback inherent in foraging exploration.

What Drives Choice: "Income" or "Return?"

With its emphasis on recent reward experience as the critical determinant of choice, our LNP model of primate choice dynamics is similar in spirit to an existing theory of matching behavior called *melioration* (Herrnstein & Vaughan, 1980; Vaughan 1982). Melioration proposes that matching results from a process in which choice continually shifts toward alternatives that yield higher return for behavioral investment, where return is defined as the average reward experienced per choice of (or time spent on) a particular option. Under steady-state conditions, melioration predicts the emergence of a behavioral equilibrium at which the returns from competing options are equal and the distribution of an animal's responses among these options is consistent with the matching law. Our present model differs from melioration with respect to both the quantity that drives behavioral change and the temporal window over which that quantity is computed.

In contrast to *return*, the driving force in our LNP model is the *income* that an animal has experienced from a particular option. To appreciate the key difference between return and income, consider the example of computing local estimates of these two quantities for the red option in our task. On the one hand, to calculate local return for the red option, we might tally the number of rewards delivered on the red option over the last, say, 10 choices to that option. To calculate income, on the other hand, we would tally the number of rewards received on red for the last 10 choices to *either* option. Tabulating rewards over choices to only one option versus choices to either option can have a strong impact on the calculated quantities. For example, if the animal has chosen the red option exclusively over the last 10 trials, then our local estimates of return and income would be numerically equal. If, however, the animal then makes a spate of choices to green, red income will decay as the red rewards are diluted by green choices. In contrast, red return will remain completely unchanged because only red choices are considered in the calculation. Thus, whereas green choices cause the recent income experienced from the red option to decay, they have no effect on the return experienced from red.

Gallistel and colleagues have highlighted fundamental inconsistencies between a return-driven choice process and the statistics of matching behavior (Gallistel & Gibbon, 2000; Gallistel et al., 2001; Mark & Gallistel, 1994). Return-based models predict that the probability of terminating a run of successive choices of a particular option ought to increase with each unrewarded choice. However, the exponential distribution of run lengths commonly observed in matching studies (Gallistel et al., 2001; Heyman, 1979, 1982; Sugrue et al., 2004) suggests that this probability is constant as a function of time. Consistent with these studies, we find that return-based models perform exceedingly poorly at both predicting and generating behavior in our matching task (data not shown). Thus, in the context of matching behavior, experimental data from several laboratories seem at odds with the prevailing paradigm in the field of reinforcement learning in which return is the dominant driver of choice behavior (Sutton & Barto, 1998).

Like our LNP model, melioration theory emphasizes the primacy of local processes in determining choice as opposed to more global processes such as maximization. Melioration, however, is vague about the exact temporal window over which average returns should be computed, making it difficult to implement as a mechanistic model that can predict or generate actual choice behavior (Baum & Aparicio, 1999). Our LNP model is considerably more specific regarding the temporal integration process, specifying a precise functional form through which recent experience influences current choice. It will be interesting to test whether our model can account for behavior in settings in which melioration has been demonstrably superior to models based upon global maximization in describing human and animal behavior (Herrnstein, Loewenstein, Prelec, & Vaughan, 1993; Vaughan, 1981; Vaughan & Herrnstein, 1987). In fact, there are some experimental findings that appear initially challenging to melioration, but may be more easily understood under income-based models like our own (e.g., Buckner, Green, & Myerson, 1993).

The Quest for Neural Correlates

Rich behavioral models can offer substantial insight into the strategies adopted by animals even in relatively complex behavioral tasks. Such models are interesting to neurophysiologists, however, because of their potential implications for the neural mechanisms that underlie behavior. The LNP model advanced here, for example, raises several obvious questions: Where in the brain are the reward integrators described in the L-stage? Do separate neural subsystems mediate the fast and slow aspects of reward evaluation? How are the differencing and sigmoidal operations for the N-stage implemented? Where is the random number generator for the P-stage? These questions all presuppose, however, that behavioral models like ours not only provide insight into behavioral strategies—the purpose for which they were developed—but also

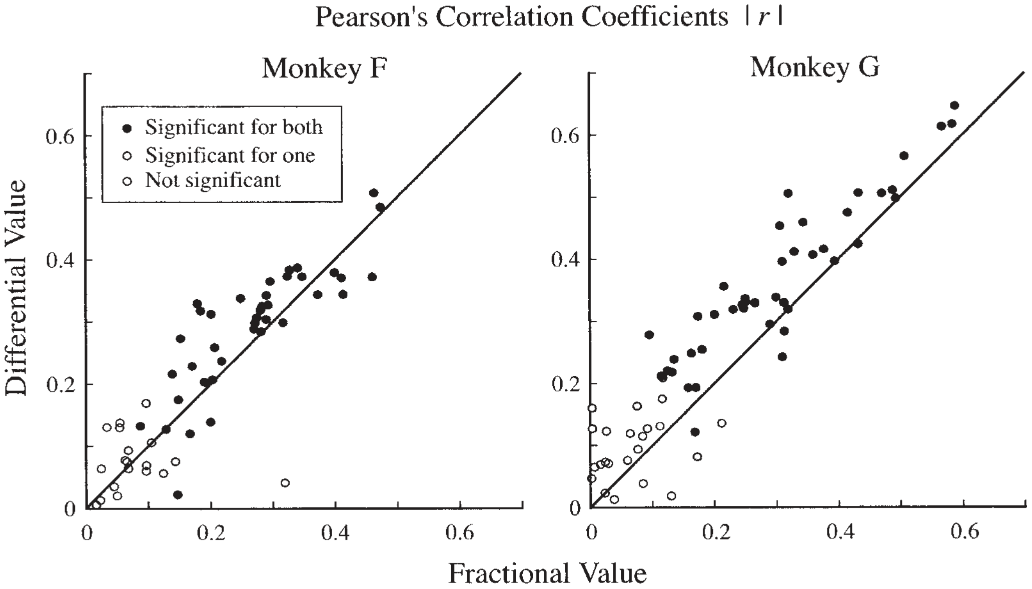


Fig. 12. Regressions of neural data onto two different value metrics. For each of 62 neurons recorded by Sugrue, Corrado, and Newsome (2004) in the same 2 animals whose behavioral data are presented here, we plot the unsigned Pearson's correlation coefficient relating each of two putative value metrics to changes in neural firing rate preceding choices both into and out of the cell's response field. Fractional value, the value metric in the original study, is shown on the abscissa and differential value, the hidden variable in the LNP model presented here, is shown on the ordinate. Regressions that were statistically significant ($p < 0.05$) for both decision variables are plotted as filled circles, those that were significant for one but not both are plotted as shaded circles, and those that were not significant for either are shown as open circles.

describe the underlying neural mechanisms that ultimately mediate the behavior. While a literal implementation of our model is one way that the brain could produce matching behavior, there are alternative models that might describe the behavioral data well but have very different internal mechanisms. Thus the success of our behavioral model *does not* necessarily imply that its components will have direct neural correlates. Therefore, we must be cautious in making any leap from behavioral description to neural mechanism.

Despite these caveats, there is reason for optimism. Models such as ours offer precise, quantitative metrics of value and choice that neurophysiologists can use to probe underlying neural systems. Indeed, evidence that neural systems might actually compute some of these metrics has emerged from an initial set of neurophysiological studies of monkeys engaged in free-choice tasks (Barraclough, Conroy, & Lee, 2004, Dorris & Glimcher 2004, Sugrue et al., 2004; see Sugrue, Corrado, & Newsome, 2005, for review). In our own prior work, for example, we found that single neurons in the lateral intraparietal area (LIP) of the primate cortex carried signals that correlated with the scalar value metric, *fractional value*, which emerged from a simple model of matching behavior developed in that earlier study. In this paper, we have developed a better behavioral model, but how does this model fare in explaining neural data? To answer this question, we reanalyzed the same neurophysiological data using differential value, the scalar value metric operationalized by our more sophisticated LNP model. Figure 12 is a scatterplot depicting how data from each of 62 LIP neurons recorded in our 2 monkeys correlate with these two scalar value metrics (details of the experimental method, analysis, and results pertaining to these neurophysiological data can be found in Sugrue et al., 2004). Both metrics generate similar numbers of significant regressions in neural data from both animals; however, correlation coefficients tend to be higher for the scalar value metric generated by our richer LNP model than for the original leaky matching rule. Thus the model presented here provides a better account of both the behavioral *and* the neural data. Clearly, we should remain wary of premature attempts to identify particular brain areas or neural computations with specific model stages. Nonetheless, initial results support guarded optimism that the operational metrics provided by rich behavioral models will prove useful for exploring the neural systems that support such complex cognitive functions as valuation and choice.

REFERENCES

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*, 404–410.

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*, 231–242.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, *32*, 269–281.

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

Baum, W. M. (1982). Choice, changeover, and travel. *Journal of the Experimental Analysis of Behavior*, *38*, 35–49.

Baum, W. M., & Aparicio, C. F. (1999). Optimality and concurrent variable-interval variable-ratio schedules. *Journal of the Experimental Analysis of Behavior*, *71*, 75–89.

Baum, W. M., & Davison, M. (2004). Choice in a variable environment: Visit patterns in the dynamics of choice. *Journal of the Experimental Analysis of Behavior*, *81*, 85–127.

Buckner, R. L., Green, L., & Myerson J. (1993). Short-term and long-term effects of reinforcers on choice. *Journal of the Experimental Analysis of Behavior*, *59*, 293–307.

Bussgang, J. J. (1975). Cross-correlation functions of amplitude-distorted Gaussian inputs. In A. H. Haddad (Ed.), *Nonlinear systems*. Stroudsburg, PA: Dowdon, Hutchinson & Ross.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network*, *12*, 199–213.

Chung, S. H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, *10*, 67–74.

Davison, M., & Baum, W. M. (2000). Choice in a variable environment: Every reinforcer counts. *Journal of the Experimental Analysis of Behavior*, *74*, 1–24.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*, 365–378.

Evarts, E. V. (1966). A technique for recording activity of subcortical neurons in moving animals. *Electroencephalography and Clinical Neurophysiology*, *24*, 83–86.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference a critical review. *Journal of Economic Literature*, *XL*, 351–401.

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

Gallistel, C. R., Mark, T. A., King, A., & Latham, P. E. (2001). The rat approximates an ideal detector of changes in rates or reward: Implications for the law of effect. *Journal of the Experimental Psychology: Animal Behavior Processes*, *27*, 354–372.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*, 769–792.

Green, L., Myerson, J., & McFadden, E. (1997). Rate of temporal discounting decreases with amount of reward. *Memory & Cognition*, *25*, 715–723.

Hays, A. V., Richmond, B. J., & Optican, L. M. (1982). A UNIX-based multiple process system for real-time data acquisition and control. *WESCON Conference Proceedings*, *2*, 1–10.

Herrnstein, R. J. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267–272.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*, 243–266.

Herrnstein, R. J., Loewenstein, G. F., Prelec, D., & Vaughan, W., Jr. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making*, *6*, 149–185.

Herrnstein, R. J., & Vaughan, W., Jr. (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. L. Commons, R. J. Herrnstein, & H. Rachlin (Eds.), *Quantitative analyses of behavior, Vol. II: Matching and maximizing accounts* (pp. 459–490). Cambridge, MA: Ballinger.

Horner, J. M., Staddon, J. E. R., & Lozano, K. K. (1997). Integration of reinforcement effects over time. *Animal Learning & Behavior*, *25*, 84–98.

Horwitz, G. D., Chichilnisky, E. J., & Albright, T. D. (2005). Blue-yellow signals are enhanced by spatiotemporal luminance contrast in macaque V1. *Journal of Neurophysiology*, *93*, 2263–2278.

Houston, A. I., & McNamara, J. (1981). How to maximize reward rate on two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior*, *35*, 367–396.

Hunter, I., & Davison, M. (1985). Determination of a behavioral transfer function: White-noise analysis of session-to-session response-ratio dynamics on concurrent VI VI schedules. *Journal of the Experimental Analysis of Behavior*, *43*, 43–59.

JEAB. (1993). *The 30 most cited articles from JEAB*. Retrieved from http://seab.envmed.rochester.edu/society/history/jeab_highly_cited.shtml

Judge, S. J., Richmond, B. J., & Chu, F. C. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research*, *20*, 535–538.

Kay, S. M. (1993). *Fundamentals of statistical signal processing.* Upper Saddle River, NJ: Prentice Hall.

Killeen, P. R. (1981). Averaging theory. In C. M. Bradshaw & E. Szabadi (Eds.), *Quantification of steady-state operant behaviour* (pp. 21–34). Amsterdam: Elsevier.

Killeen, P. R. (1994). Mathematical principles of reinforcement: Based on the correlation of behavior with incentives in short-term memory. *Behavioral and Brain Sciences*, *17*, 105–172.

Kirby, K. N., & Maraković, N. N. (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organization Behavior and Human Decision Processes*, *64*, 22–30.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*, 555–579.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *Economic Journal*, *97*, 666–684.

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

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior: Vol. 5. The effect of delay and of intervening events on reinforcement value* (pp. 55–73). Hillsdale, NJ: Erlbaum.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004, October 15). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*, 503–507.

McDowell, J. J (1980). An analytic comparison of Herrnstein's equations and a multivariate rate equation. *Journal of the Experimental Analysis of Behavior*, *33*, 397–408.

McDowell, J. J, Bass, R., & Kessel, R. (1983). Variable-interval rate equations and reinforcement and response distributions. *Psychological Review*, *90*, 364–375.

McDowell, J. J, & Kessel, R. (1979). A multivariate rate equation for variable-interval performance. *Journal of the Experimental Analysis of Behavior*, *31*, 267–283.

Palya, W. L., Walter, D., Kessel, R., & Lucke, R. (1996). Investigating behavioral dynamics with a fixed-time extinction schedule and linear analysis. *Journal of the Experimental Analysis of Behavior*, *66*, 391–409.

Palya, W. L., Walter, D., Kessel, R., & Lucke, R. (2002). Linear modeling of steady-state behavioral dynamics. *Journal of the Experimental Analysis of Behavior*, *77*, 3–27.

Rachlin, H., Battalio, R., Kagel, J., & Green, L. (1981). Maximization theory in behavioral psychology. *Behavioral and Brain Sciences*, *4*, 371–388.

Shahan, T. A., & Lattal, K. A. (1998). On the functions of the changeover delay. *Journal of the Experimental Analysis of Behavior*, *69*, 141–160.

Simoncelli, E. P., Paninski, L., Pillow, J. W., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences III* (3rd ed., pp. 327–338). Cambridge, MA: MIT Press.

Southwick, C. H., & Siddiqi, M. F. (1985). Rhesus monkey's fall from grace. *Natural History*, *94*(2), 63–70.

Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*, 165–180.

Stubbs, D. A., Pliskoff, S. S., & Reid, H. M. (1977). Concurrent schedules: A quantitative relation between changeover behavior and its consequences. *Journal of the Experimental Analysis of Behavior*, *27*, 85–96.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004, June 18). Matching behavior and the representation of value in parietal cortex. *Science*, *304*, 1782–1787.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews in Neuroscience*, *6*, 363–375.

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

Vaughan, W., Jr. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior*, *36*, 141–149.

Vaughan, W., Jr. (1982). Choice and the Rescorla-Wagner model. In M. L. Commons, R. J. Herrnstein, & H. Rachlin (Eds.), *Quantitative analyses of behavior. Vol. II: Matching and maximizing accounts* (pp. 263–279). Cambridge, MA: Ballinger.

Vaughan, W., Jr., & Herrnstein, R. J. (1987). Stability, melioration, and natural selection. In L. Green, & J. H. Kagel (Eds.), *Advances in behavioral economics.* (Vol. 1, pp. 185–215). Norwood, NJ: Ablex.

Received April 1, 2005
Final acceptance November 14, 2005

APPENDIX

The Soundness of Model Estimation

Here we address two concerns regarding the soundness of our LNP model estimation procedure. First, we assess the stability of our L-stage estimates: if we estimate a linear filter iteratively, do we obtain the same answer repeatedly? Second, we would like to explore our capacity to recover arbitrary L-stage filters: can we recover an arbitrary filter of a known shape, or is this procedure biased to recover filters of the shape we extracted from the real data?

To address the question of stability, we can leverage our simulated task environment. The top of Figure A1A diagrams our approach. We (a) estimate an LNP model from real behavioral data, (b) generate artificial data in simulated play using the estimated LNP, and then (c) reestimate an LNP model from these artificial data. If our LNP model estimation procedure is stable, we should recover the same LNP model that generated the artificial data when we perform the repeat estimation. The solid line in the bottom of Figure A1A shows the fit to the recovered kernel for Monkey F—this is identical to the black fit line shown in Figure 4A of the main text. The parameters from this fit describe the LNP model that we use to generate artificial choice data in simulation. The open data points in Figure A1A then show the filter recovered when we reapplied our estimation procedure, extracting the filter from the artificial data as though it were real animal data. The repeated-estimation filter closely matches the double exponential filter that generated the artificial data from which it was extracted. This shows that our estimation procedure is stable under these conditions, recovering the same filter when applied iteratively.

It still might be that the estimation procedure is strongly biased to recover compound exponential kernels like the ones we extracted from our data set, and that while those estimates are stable under simulation and reestimation, they do not reflect the kernel that generated the original data. We address this possibility directly by simulating choice data using an arbitrary L-stage kernel, and attempting reconstruction. Figure A1B diagrams this procedure and shows the results. The solid line in the bottom panel shows an arbitrary kernel that was used to generate synthetic choice data over the same sequence of blocks that our animals were exposed to. The open points show the recovered filter estimate. The fidelity of the reconstruction is quite high. Subtle features like differences in the height of the steps along the length of the filter can be made out. Even sharp transitions such as those of the high-frequency spikes at the tail of the filter are recovered quite well. Larger data sets would be required to recover these features more crisply. The greatest deviation comes at trials close to zero lag where the reconstruction fails to capture the downward corner of the filter's triangular ramp, but these small deviations are far from the distortion or bias that might cause us to confuse an arbitrary filter with one having a double exponential form similar to that recovered from the real data.

Taken together, these exercises attest to the integrity of our LNP model estimation procedure and to its usefulness in extracting behavioral models even in situations where many of the assumptions that guarantee its performance are violated.

The Computation of Average-Likelihood

We are given a binary data set, d, of length N. A probabilistic model, p, predicts that $d(i)$ will take the value 1 with probability $p(i)$. We would like to compute how likely the data d are under this model p.

The likelihood of d given p can be computed directly, by multiplying by $p(i)$ on all instances where $d(i) = 1$ and multiplying by $1 - p(i)$ on all instances where $d(i) = 0$:

$$L(d|p) = \prod_{\substack{\text{for } i \text{ where} \\ d(i)=1}} p(i) \times \prod_{\substack{\text{for } i \text{ where} \\ d(i)=0}} (1 - p(i)). \qquad \text{(A1)}$$

This formulation is very cumbersome, as well as numerically difficult to handle, and thus it is often preferred to compute the log-likelihood. By taking the log of Equation A1, we can convert the products to sums:

$$\log L(d|p) = \sum_{\substack{\text{for } i \text{ where} \\ d(i) = 1}} \log[p(i)] + \sum_{\substack{\text{for } i \text{ where} \\ d(i) = 0}} \log[1 - p(i)]. \qquad \text{(A2)}$$

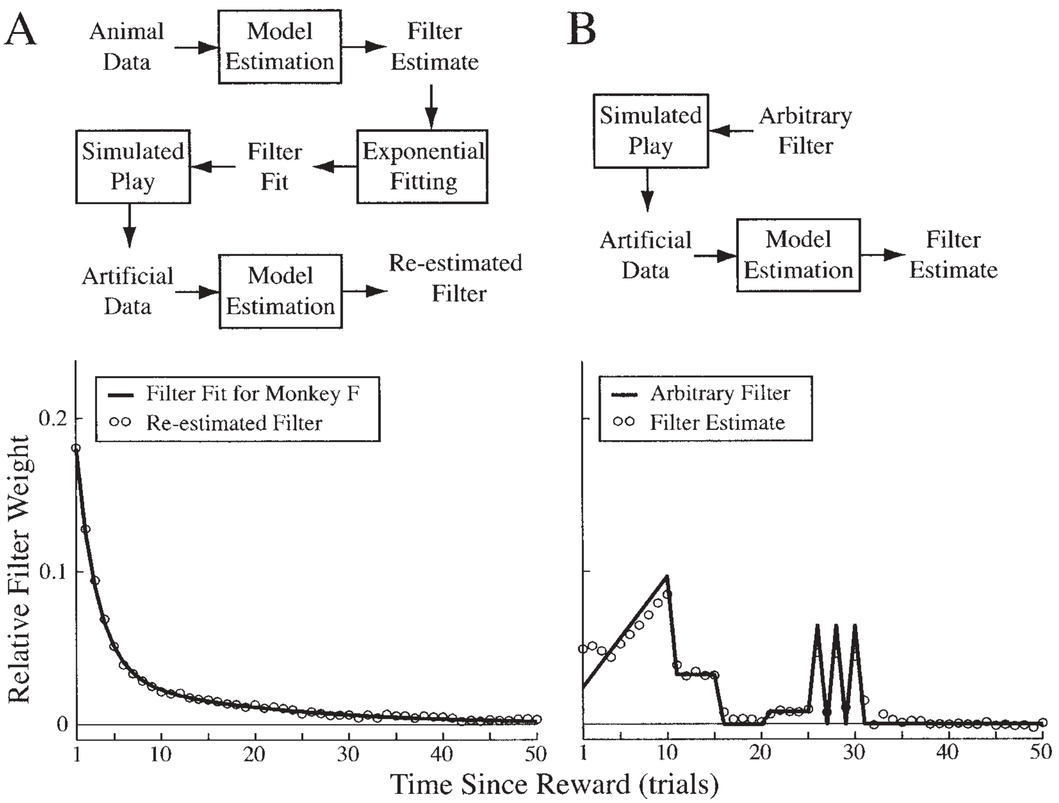


Fig. A1. (A) Kernel estimation stability. The top half of the panel diagrams the procedure used. Starting from raw behavioral data obtained from a particular animal, a linear filter was reconstructed using Wiener kernel analysis. This raw kernel was then fit with a double exponential, shown by the solid line (bottom panel). Artificial data were generated in simulation using this fit kernel, and then a new kernel was reconstructed from these artificial data. This repeated estimate of the kernel is shown in the open data points (bottom panel). (B) Arbitrary kernel estimation. An arbitrary linear filter (the solid line in the bottom panel) was used to generate synthetic data in simulation as diagrammed in the top half of the panel. These data were generated using the same block sequences, and were of the same size, as those presented to the monkeys in actual behavioral experiments. The open data points (bottom panel) show the recovered estimate of the underlying filter based on these artificial data. Recovered kernels in both A and B fit the input kernels quite well.

This measure accumulates a sum from $i = 1$ through N, adding $\log[p(i)]$ if $d(i) = 1$ and $\log[1 - p(i)]$ if $d(i) = 0$. This is much simpler to compute, but still has an undesirable dependence on N, the length of the data set.

We would prefer a measure that is independent of the size of the data set, in essence a "per trial" metric of the likelihood of the data given the model. To meet this need, we define the average-log-likelihood simply by normalizing Equation A2 by N, the length of the data set:

$$\text{Avg}\log L(d|p) = \frac{1}{N}\left[\sum_{\substack{\text{for } i \text{ where}\\ d(i)=1}} \log p(i) + \sum_{\substack{\text{for } i \text{ where}\\ d(i)=0}} \log[1-p(i)]\right]. \quad \text{(A3)}$$

This metric is easy to compute and has the desired independence from the length of the data set. We can make the numerical values easier to interpret by undoing the logarithmic transformation, simply by exponentiating Equation A3:

$$\operatorname{Avg} L(d|p) = \exp\left\{ \frac{1}{N} \left[\sum_{\substack{\text{for } i \text{ where} \\ d(i)=1}} \log p(i) + \sum_{\substack{\text{for } i \text{ where} \\ d(i)=0}} \log[1-p(i)] \right] \right\}. \quad \text{(A4)}$$

Now an average-likelihood value of 1.0 is easily understood as perfect prediction of every trial, whereas a value of 0.0 corresponds to exactly wrong predictions, and a value of 0.5 indicates that the model performs at chance levels averaged across the entire data set.