Steyvers, M., Lee, M.D., & Wagenmakers, E.J. (in press). A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology.

A Bayesian Analysis of Human Decision-Making on Bandit Problems

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

Eric-Jan Wagenmakers
Department of Psychology
University of Amsterdam

Abstract

The bandit problem is a dynamic decision-making task that is simply described, well-suited to controlled laboratory study, and representative of a broad class of real-world problems. In bandit problems, people must choose between a set of alternatives, each with different unknown reward rates, to maximize the total reward they receive over a fixed number of trials. A key feature of the task is that it challenges people to balance the exploration of unfamiliar choices with the exploitation of familiar ones. We use a Bayesian model of optimal decision-making on the task, in which how people balance exploration with exploitation depends on their assumptions about the distribution of reward rates. We also use Bayesian model selection measures that assesses how well people adhere to an optimal decision process, compared to simpler heuristic decision strategies. Using these models, we make inferences about the decision-making of 451 participants who completed a set of bandit problems, and relate various measures of their performance to other psychological variables, including psychometric assessments of cognitive abilities and personality traits. We find clear evidence of individual differences in the way the participants made decisions on the bandit problems, and some interesting correlations with measures of general intelligence.

Introduction

Bandit Problems

Imagine you are staying in an unfamiliar city for a few weeks, and are consigned to eat alone each evening. There are a number Chinese restaurants, all with essentially the same menu, and all within a short walk of your hotel. Each menu is cheap enough (or your expense account is large enough) that whether the meal is 'good' or 'bad' acts as the sole criterion for choosing one restaurant over the other.

Over the course of your stay, a natural dining goal would be to maximize the number of good meals. In the first few days, pursuing this goal might involve trying a number of the restaurants. If the meal on the first night was bad, it seems unlikely you would re-visit that restaurant on the second night.

Towards the end of your stay, however, it becomes increasingly likely you will visit the same reasonably good restaurant repeatedly, even if it does not produce good meals every night. There is less incentive to explore options about which you are less certain, and more incentive to exploit options which are reasonably good, and about which you are more certain. Because of the limited nature of your stay, how you search the environment of restaurants is likely to move from an initial *exploration* phase to a mature *exploitation* phase.

Another source of information that will affect your decision-making is any potential knowledge you have about the quality of Chinese restaurants in the city. If the city has a reputation for having many excellent Chinese restaurants, then a smaller number of bad meals will encourage the trying of alternatives. On the other hand, if you believe the city does not have many good Chinese restaurants, a relatively modest success rate at one place might encourage repeat dining.

It seems clear that your decision-making in trying to maximize the number of good meals you have is a non-trivial optimization problem. Choices must be sensitive to previous dining outcomes, how many days of your stay remain, what you know about the relevant base-rates for good and bad meals, and the interaction of all these factors.

This real world decision-making scenario has the same essential characteristics as a formal mathematical optimization problem known as the 'bandit' problem[1], which, following its original description by Robbins (1952), has been studied extensively in the machine learning and statistics literatures (e.g., Berry, 1972; Berry & Fristedt, 1985; Brezzi & Lai, 2002; Gittins, 1979, 1989; Kaebling, Littman, & Moore, 1996; Macready & Wolpert, 1998; Sutton & Barto, 1988). In a general N-armed bandit problem, there is a set of N bandits, each having some fixed but unknown rate of reward. On each trial, a decision-maker must choose a bandit, after which they receive feedback as to whether or not one unit of probabilistically determined

[1] The name 'bandit' problem comes from an analogy with multiple-armed gaming machines, sometimes known as slot machines or bandit machines. From this analogy, a sequence of trials on the bandit problem is often called a 'game'.

reward was attained. The decision-maker's task is, over a set of K trials, to use this feedback to make a sequence of bandit choices that maximizes their reward.

Why Study Bandit Problems?

We believe that bandit problems provide an interesting and useful task for the study of human capabilities in decision-making and problem solving. They provide a challenging task, similar to many real-world problems that is nevertheless simple to understand. They require people to search their environment in intelligent ways to make decisions, exploring uncertain alternatives and exploiting familiar ones. The ability to search effectively, striking the right balance between exploration and exploitation, is a basic requirement for successful decision-making (Gigerenzer & Todd, 1999). Bandit problems also have a known optimal solution process to which human decision-making can be compared. Being able to make the distinction between 'outcome optimal' and 'process optimal' decision-making is useful, because adherence to the process can always be achieved, but attainment of the reward is inherently stochastic. This property can make process optimality a less noisy measure of decision-making performance than the decision optimality (Lee, 2006).

Early studies of human performance on bandit problems used models and experimental manipulations motivated by theories of operant conditioning (e.g., Brand, Wood, & Sakoda, 1956; Brand, Sakoda, & Woods, 1957). Later studies were informed by economic theories, leading to a focus on deviations from rationality in human decision-making (e.g., Anderson, 2001; Banks, Olson, & Porter, 1997; Horowitz, 1973; Meyer & Shi, 1995). Most of these economic studies focused on two-choice bandit problems, but considered variants on the basic bandit problem we have defined. In particular, they sometimes fixed the reward rate of one of the two choices, or motivated by the economic concept of an 'infinite horizon' (i.e., the potential of an arbitrarily long sequence of decision-making between the alternatives), considered bandit problems without a fixed number of trials, but instead introduced a small probability of terminating the problem after any given trial. Human performance on the bandit problem has also been a recent focus of interest in cognitive neuroscience (e.g., Cohen, McClure, & Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

One issue that deserves more systematic study involves individual differences in solving bandit problems. The ability to solve problems involving the maximization (or minimization) of some criterion under multiple interacting constraints is generally regarded as a basic expression of intelligence. Accordingly, how well people solve optimization problems like bandit problems may provide an interesting window onto differences in basic human cognitive abilities. A good recent example of this line of inquiry is provided by Burns, Lee, and Vickers (2006), who explored the relationship between human performance on optimal stopping problems known as secretary problems (Gilbert & Mosteller, 1966) and standard psychometric measures of cognitive abilities. These authors demonstrated that performance on the Secretary Problem loaded on fluid intelligence (G_f), with performance on the problem also being shown to load approximately 0.4 on a general ability factor, (g). This g-loading was comparable to that of the Digit Symbol task from the Wechsler Adult Intelligence Scale. In a similar spirit, it seems worth examining whether people's abilities to follow optimal decision processes in bandit problems differ, and whether any differences are related to psychometric measures of cognitive abilities.

Performance on bandit problems also seems to have natural links to personality traits that control risk behavior. Too much exploration in solving a bandit problem could be regarded as a form of risk seeking behavior, while too much exploitation could be regarded as risk averse behavior. Recently, the risk propensity of clinical and normal populations has been a research focus in neuropsychology. In particular, clinical populations with damage to the ventromedial prefrontal cortex, cocaine abusers, and patients with Aspberger's syndrome often take decisions that are risk seeking relative to a normative analysis (Bechara, Damasio, Damasio, & Anderson, 1994; Yechiam, Busemeyer, Stout, & Bechara, 2005). In a laboratory setting, risk-taking behavior is often quantified using tasks such as the Iowa Gambling Task (e.g., Bechara et al., 1994; Busemeyer & Stout, 2002; Wood, Busemeyer, Koling, Cox, & Davis, 2005), which can be conceived as a special type of bandit problem, and the Balloon Analog Risk Task (e.g., Wallsten, Pleskac, & Lejuez, 2005). The classic bandit problems we consider provide another task for continuing and expanding the study of individual differences in risk-related behavior.

Overview

In this paper, we study individual differences in how people balance exploration and exploitation in solving bandit problems, using a natural Bayesian extension of the optimal decision model. The basic idea is to extend the optimal decision process to operate in a range of environments, characterized by different distributions over the reward rates. The insight is that people who explore (i.e., choose alternatives about which less in known) can be regarded as making optimistic assumptions about the underlying reward rates, while people who exploit (i.e., choose alternatives about which more is known) can be regarded as making more pessimistic assumptions. Within this framework, differences in human decision-making can be explained in terms of differences in the assumptions people make about the statistical structure of their environment.

We begin by giving a formal characterization of the optimal Bayesian decision process for bandit problems, parameterized by assumptions about the distribution of underlying reward rates. We develop a method for inferring these parameters from the decisions people make solving bandit problems. These methods are then applied to data from 451 participants, for whom extensive additional cognitive ability and personality trait measures are also available. From a basic analysis, we observe that there are individual differences that make it useful to be able to measure the extent to which people adhere to the optimal decision process, rather than simpler heuristic strategies. Accordingly, we develop a method based on Bayesian model selection to compare the optimal decision model to a three alternative heuristic models, with varying degrees of psychological sophistication. When we fit the models and their parameters to human data, we find clear evidence of individual differences, both in terms of which model is used, and what the interesting parameter values are. Finally, we use our models, together with simpler experimental measures of performance, to examine how the decision-making of all 451 participants on the bandit problems relates to measures of their cognitive abilities and personality traits.

Optimal Bayesian Decision-Making for Bandit Problems

Environmental Assumptions

We assume that the way people might think about bandit problem environments can be modeled by allowing the reward rates to come from any Beta distribution. In its standard form, the Beta distribution has two parameters: a count α of 'prior successes' and a count β of 'prior failures'. A natural and useful re-parameterization is to consider $\alpha/\left(\alpha+\beta\right)$, which is a measure of the expected rate of reward, and $\alpha+\beta$, which is a measure of the strength of the prior evidence. Psychologically, $\alpha/\left(\alpha+\beta\right)$ corresponds to something like the level of 'optimism' about reward rates, while $\alpha+\beta$ corresponds to something like the 'certainty' with which that level of optimism is held.

A consequence using the flexible Beta environment is that, based on exactly the same task information, different values of α and β will lead to different optimal decisions. Optimistic assumptions about the nature of the environment, corresponding to believing that it is likely alternatives will have high reward rates, will tend to lead to optimal decisions involving exploration on largely untested options. Pessimistic assumptions about the environment, corresponding to believing that it is likely alternatives will have low reward rates, will tend to lead to optimal decision involving the exploitation of relatively well known, if not highly rewarding, alternatives. We note that this perspective on the balance of exploration and exploitation is quite different from standard reinforcement learning accounts, such as Q-learning (see Sutton & Barto, 1988), which view exploration as resulting from the occasional introduction of randomness into decision-making processes.

Figure 1 provides an example of this relationship between environmental assumptions and optimal decisions. The left panel shows the state of a game with four alternatives, after eight of ten trials have been completed. The x-axis corresponds to the number of failures for each alternative. The y-axis corresponds to the number of successes. Accordingly, the previous reward history of each alternative is represented as a point the space, labeled by that alternative's numbers. In Figure 1, the first alternative has not been chosen; the second alternative has been chosen twice, for one success and one failure; the third alternative has been chosen five times, for two

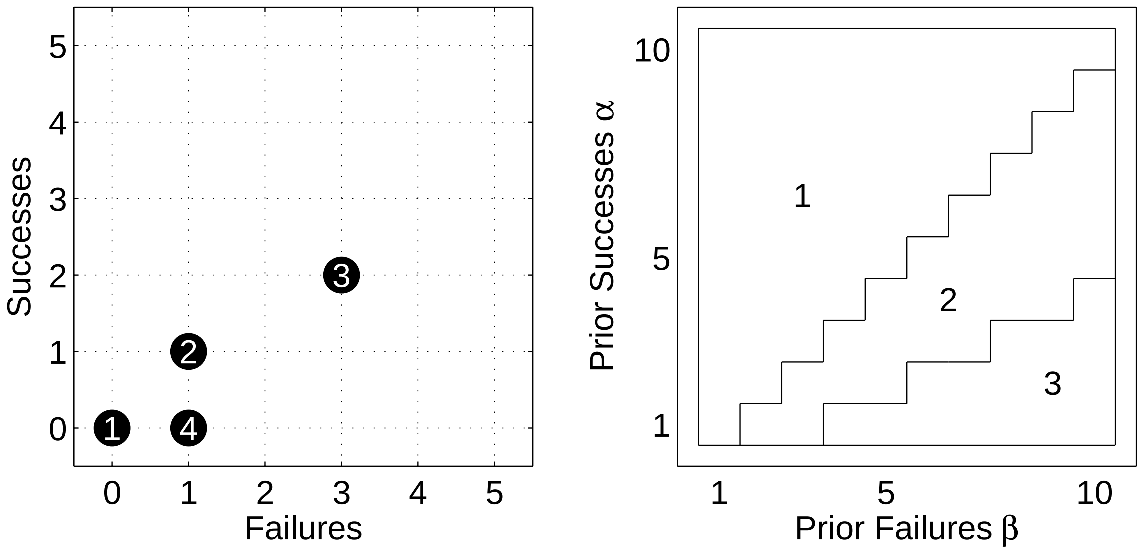

Figure 1. Optimal decision-making for the same alternatives in a four-choice ten-trial game, under different assumptions about the environment.

successes; and the fourth alternative has been chosen once, for a lone failure.

The right hand panel of Figure 1 shows the optimal decision (we explain how this is calculated in the next section), as a function of different assumptions about the nature of the environment as measured by α and β. If it is assumed that $\alpha \geq \beta$, so that there are many high reward alternatives, then the optimal choice is the previous unseen first alternative. If, on the other hand, β is much greater than α, so that their are many low reward rates, the third alternative, which has been successful the most often is the optimal choice, particularly since only two trials remain in the game. Between these extremes, the optimal choice is the second alternative, for which one success and one failure has been recorded. The fourth alternative can never be the optimal choice.

Optimal Decision-Making in Beta Environments

As we have already mentioned, for a fixed environment, the bandit problem has a known optimal solution process (see, for example Kaebling et al., 1996, p. 244). Because of the finite horizon of our bandit problems, the solution does not rely on well known Gittin's (1989) indices. Rather, the solution can be explicitly derived by recursive methods for small enough numbers of trials and alternatives. The basic idea is that, on the last trial, the alternative with the greatest expected reward should be chosen. On the second-last trial, the alternative that leads to the greatest expected total reward over the remaining two trials should be chosen, assuming the last trial is chosen optimally in the way already established. On the third-last trial, the alternative that leads to the greatest total reward for the remaining three trials should be chosen, assuming the final two choices are also optimal, and so on. By continuing backwards through the trial sequence in this way, it is possible to determine the optimal decision process for the entire problem.

Formally, our goal is to define an optimal Bayesian decision process for a bandit problem, under the assumption that the underlying reward rates are independent samples from a $\text{Beta}\big(\alpha^*, \beta^*\big)$ distribution. Denoting the reward rate for the ith alternative as θ_i^g, we can write $\theta_i^g \sim \text{Beta}\big(\alpha^*, \beta^*\big)$. The constants α^* and β^* define the true environmental distribution from which the reward rates are drawn, and should be distinguished from the α and β parameters of the decision model, which index the range of possible environments that participants could assume.

We denote the decision on the kth trial of the gth game by D_k^g, and the set of all decisions in a condition by D. On the kth trial, if the ith alternative is chosen, whether or not a reward is attained is determined as the outcome of a Bernoulli trial using the reward rate θ_i^g, so that $R_k^g \sim \text{Bernoulli}\big(\theta_{D_k^g}\big)$.

To define the optimal decision process with respect to $\text{Beta}\big(\alpha, \beta\big)$ environments, we let $S_k^g = \{s_1, f_1, \ldots, s_N, f_N\}$ be the state of the gth game after k trials, where s_i and f_i count the number of existing successes and failures for the ith alternative. This is appropriate, because for decision-making, the only relevant aspects of the game are these counts. We let $u_k^{\alpha,\beta}\left(s_1, f_1, \ldots, s_N, f_N\right)$ be the expected total additional reward to be gained by following the optimal decision process for the remaining trials. Following Kaebling et al. (1996), this can be calculated recursively, using the update equation

$$\begin{aligned} u_k^{\alpha,\beta}\left(s_1, f_1, \ldots, s_N, f_N\right) &= \max_i E\big[\text{Future reward for choosing } i\text{th alternative,} \\ &\qquad\qquad \text{if all subsequent decisions are optimal}\big] \\ &= \max_i \Bigg[\frac{s_i + \alpha}{s_i + f_i + \alpha + \beta} u_{k+1}^{\alpha,\beta}\left(s_1, f_1, \ldots, s_{i+1}, f_i, \ldots, s_N, f_N\right) + \\ &\qquad \frac{f_i + \beta}{s_i + f_i + \alpha + \beta} u_{k+1}^{\alpha,\beta}\left(s_1, f_1, \ldots, s_i, f_{i+1}, \ldots, s_N, f_N\right)\Bigg]. \end{aligned} \tag{1}$$

In this equation, the term $\left(s_i + \alpha\right) / \left(s_i + f_i + \alpha + \beta\right)$ is the probability of a success for the ith alternative in a $\text{Beta}\big(\alpha, \beta\big)$ environment, and $\left(f_i + \beta\right) / \left(s_i + f_i + \alpha + \beta\right)$ is the probability of a failure. The expected additional reward for the last trial $u_K^{\alpha,\beta} = 0$, which completes the recursive definition.[2]

The optimal choice, for a specific environment given by α and β, is simply one

[2] The recursive definition to compute expected reward can be programmed efficiently using dynamic programming techniques that exploit symmetries in the game states. For example, for $N = 4$, and $K = 15$ (the values we will use in our experiment), there are only 35,876 unique game states that need to be explicity represented to calculate the expected rewards of any game.

that maximizes the total expected reward given by Equation 1, so that

$$p(D_k^g = i \mid S_k^g, \alpha, \beta) = \begin{cases} 1 & \text{if the } i\text{th alternative maximizes the total expected reward,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Allowing for Suboptimal Decisions

One practical difficulty with Equation 2 is that zero likelihood is given to sub-optimal decisions. From the perspective of making inferences about human decision-making, this is undesirable, because it means a person's behavior can be completely consistent with some (α, β) combination for all but one decision, but zero posterior density will be given to that combination as a result of the one sub-optimal decision.

To address this problem, we introduce an additional responding parameter w. This parameter is a rate that controls the level of determinism in decision-making, or, equivalently, the "accuracy of execution" of the deterministic optimal decision rule. Under this conception, the responding parameter is not part of the optimal decision-making process, but, rather, characterizes the inevitable noise in human behavioral data. For this reason, we continue to implement the optimal decision-process in its traditional form, although we note that if the responding parameter were to be included, this optimal process would have to be augmented.

The basic idea in our approach is that, for high rates of w, the optimal rule is followed exactly, but as the rate w decreases, occasionally a non-optimal decision is made. There are many ways this idea can be implemented formally, including by soft-max or the Luce (1963) choice rule, or using entropification methods (Grünwald, 1998; Lee, 2004). We chose a simple implementation, in which

$$p(D_k^g = i \mid S_k^g, \alpha, \beta, w, M_{\text{opt}}) = \begin{cases} w/N_{\max} & \text{if the } i\text{th alternative maximizes total expecte} \\ (1-w)\,/\,(N - N_{\max}) & \text{otherwise,} \end{cases} \tag{3}$$

where N is the number of alternatives, and $N_{\max}$ is the number of alternatives that maximize reward for a given combination of α and β.

Inference Using Graphical Modeling

A graphical model representation (see Jordan, 2004 and Koller, Friedman, Getoor, & Taskar, 2007 for statistical introductions, and Griffiths, Kemp, & Tenenbaum, 2008 and Lee, 2008 for psychological introductions) of this Bayesian model is shown in Figure 2. The variables of interest are represented as nodes in a graph. Observed variables, including the decisions of participants (D), the rewards obtained from choices (R), and the states of the game (S) that follows from these decisions and rewards have shaded nodes. Because the states are determined by the sequence of decisions and rewards, they are deterministic, and have double-lined border. All of the other variables are stochastic. Unobserved variables, including the true nature of the environment (α^*, β^*), the reward rates (θ) sampled from the environment, the

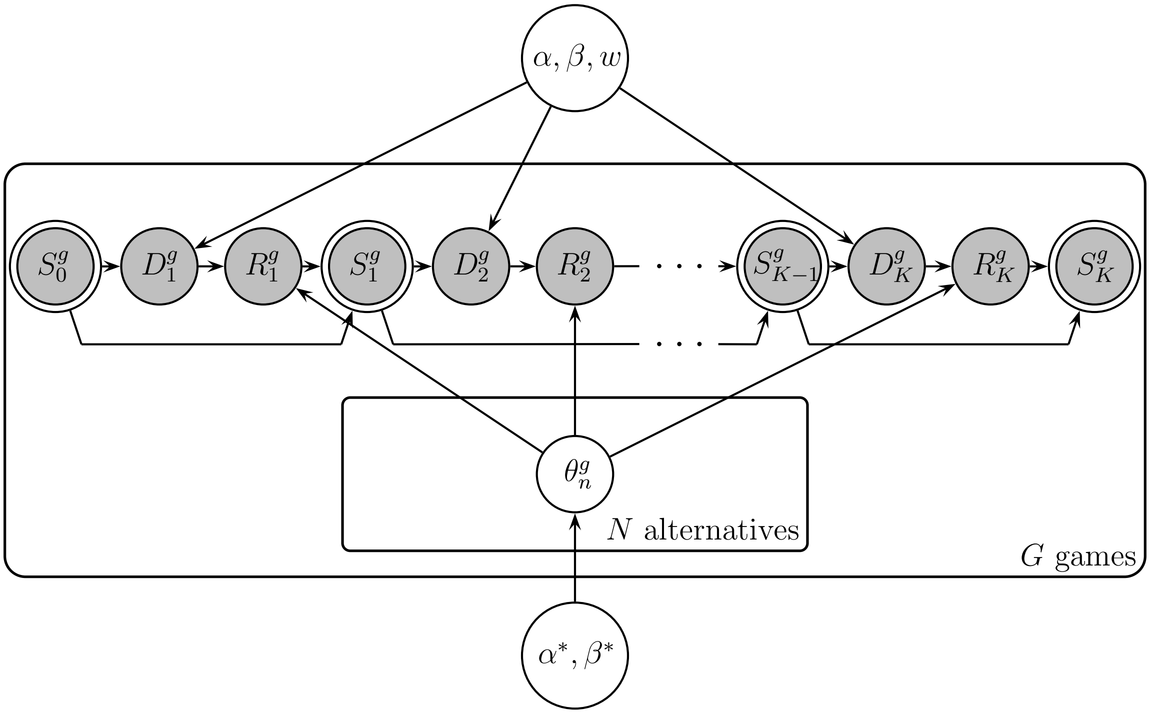

Figure 2. Graphical model representation of Bayesian optimal decision-making in Beta environments.

structure assumed by the participant (α, β), and the determinism of their responding (w) are unshaded. Directed arrows indicate the dependencies between variables in the model.

The encompassing plates indicate independent replications, so that there are N alternatives within a game, and G independent games for an experimental session. To extend our notation to cover separate games, we use θ_i^g to denote the reward rate for the ith alternative in the gth game, S_k^g for the state of the gth game after the kth trial, D_k^g for the kth decision in the gth game, and R_k^g for the reward obtained on the kth trial of the gth game.

To undertake inference using this graphical model, the probability of a particular α, β combination, based on a series of decisions D, is given by

$$\begin{aligned} p\big(\alpha, \beta, w \mid D\big) &\propto p\big(D \mid \alpha, \beta, w\big) p\big(\alpha, \beta, w\big) \\ &= \left[\prod_{g=1}^{G}\prod_{k=1}^{K} p\big(D_k^g \mid S_k^g, \alpha, \beta, w\big)\right] p\big(\alpha, \beta\big) p\big(w\big). \end{aligned} \tag{4}$$

The conditional probabilities $p\big(D_k \mid S_k, \alpha, \beta\big)$ are tabulated according to the optimal decision process for bandit problems given in Equation 3. We simply assume a uniform prior on the responding parameter w. Determining a prior for the beliefs about the environment is less straightforward. Gelman, Carlin, Stern, and Rubin (2004, p. 128) suggest an appropriate vague prior distribution is $p(\alpha, \beta) = (\alpha+\beta)^{-\frac{5}{2}}$. Their argument is that this prior is uniform over the psychologically interpretable parameterization $\alpha/(\alpha + \beta)$ and $(\alpha + \beta)^{-1/2}$. This is an appealing result, that we adopt. It places a uniform prior over the optimism about reward rates, and favors smaller levels of certainties or prior evidence, without being too prescriptive.

Experiment

Subjects

A total of 451 participants completed a series of bandit problems, as well as battery of other psychological tests, as part of 'testweek' at the University of Amsterdam.

Method

Each participant completed 20 bandit problems in sequence. All of the problems had 4 alternatives and 15 trials, and drew the reward rates for each alternative independently from a $\text{Beta}\big(2, 2\big)$ distribution (i.e., we set $\alpha^* = \beta^* = 2$). The drawing of rates for the games was done only once, so each participant played games with the same θ values, but the order of the games was randomized.

A representation of the basic experimental interface is shown in Figure 3. The four large panels correspond to the four alternatives, each of which can be chosen on any trial by pressing the button below. Within the panel, the outcomes of previous choices are shown as count bars, with good outcomes on the left, and bad outcomes on the right. At the top of each panel, the ratio of good to bad outcomes, if defined, is shown. The top of the interface provides the count of the total number of good outcomes to the current point in the game, the position of the current trial within the sequence, and the position of the current game within the overall set.[3]

Thus, in Figure 3, the first game of 20 involving four alternatives over 15 trials is being played. Seven trials in this game have been completed, with the first, second and third alternatives having been chosen once, three, and three times respectively. Choosing the first alternative has produced one bad outcomes; the second and third alternatives have both produced two good and one bad outcomes. The fourth alternative has not been chosen.

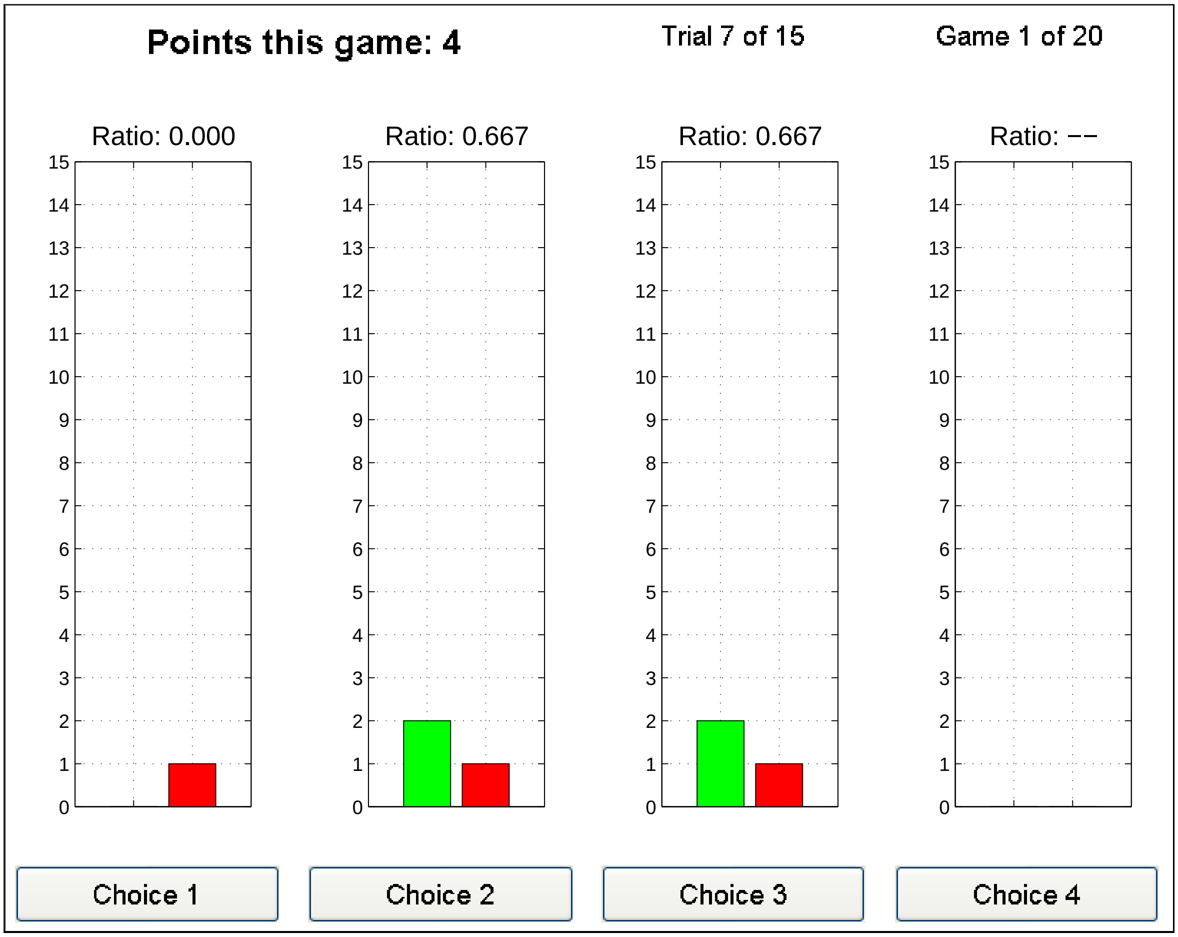

Figure 3. Basic design of the experimental interface.

Basic Results

Figure 4 shows the basic experimental results that motivate our modeling of the data. The distribution of average reward per trial is shown for all participants, and is compared to three other distributions. These other distributions come from 451 simulated participants, following different decision models. On the left is the distribution of reward achieved by simulating 'random' decisions, which means having equal probability of choosing each of the alternatives on each trial.

On the right are two optimal distributions. One corresponds to following exactly the optimal decision process. The other is a yoked optimal distribution. This is

[3] The actual software used for data collection relied on a 'fishing game' cover story, in which each choice was a pond, and either a frog was caught or was not caught on every trial. The layout was otherwise the same as Figure 3, but the more complicated graphic elements in the real interface are not well suited to journal reproduction.

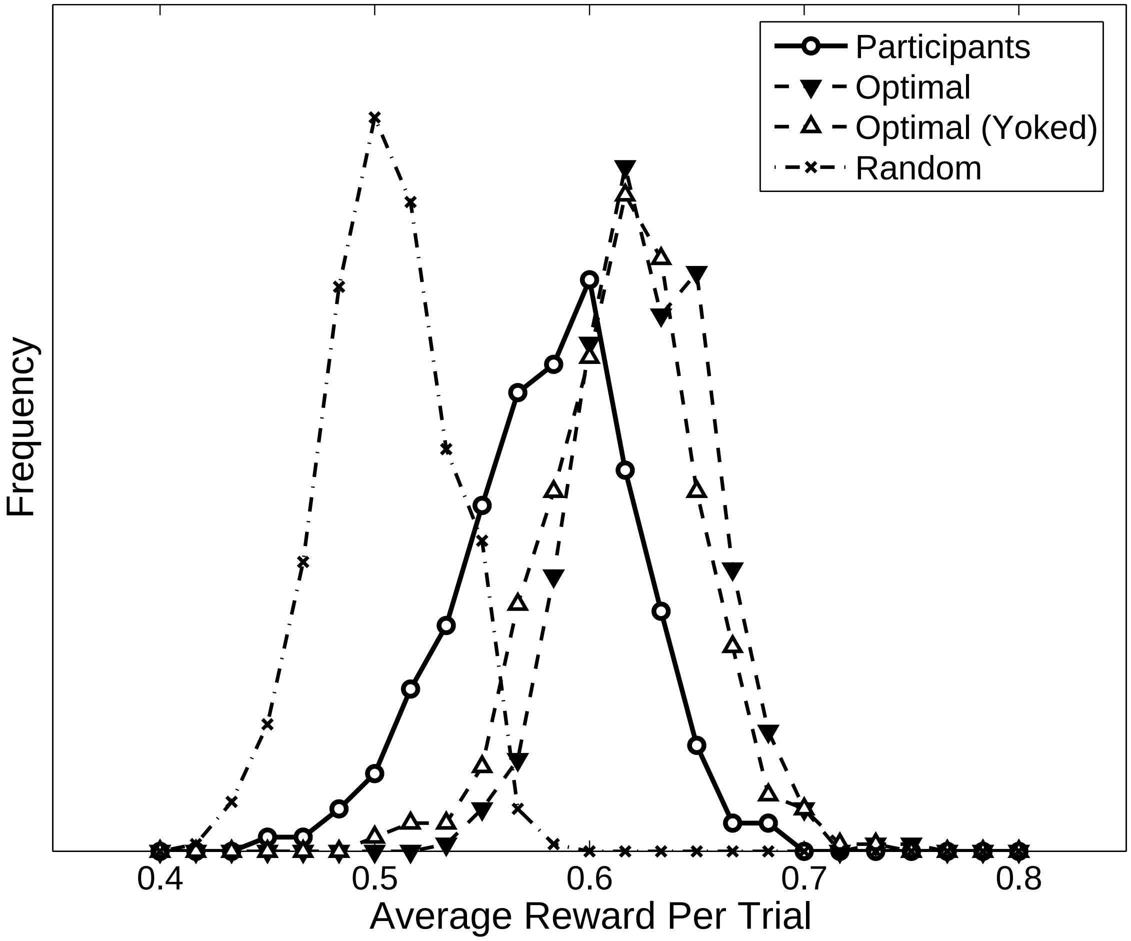

Figure 4. The distribution of average reward per trial for all participants, optimal decision-making, yoked optimal decision-making, and random decision-making.

found by taking the state of the game at each trial, calculating what the return would have been for that trial if an optimal decision was followed, and then summing the scores across trials. In this procedure, the game state is not updated with the optimal decision. Rather, the game state corresponding to the actual decision of the participants is always used. In other words, the yoked distribution corresponds to the return of the optimal decision process a trial-by-trial basis, with all game states determined by the decisions of the participants.

It is clear from Figure 4 that reward associated with the decision-making of our participants is neither as good as optimal nor as bad as random, but overlaps considerably with both. The fact that the yoked distribution is very similar to the optimal distribution suggests that it is not the case participants make one early sub-optimal decision, and cannot recover to achieve a good reward by behaving optimally from the remainder of the game. That is, the yoked optimal analysis tell us it is likely that participants make sub-optimal decisions on many trials, rather than crucial bad decisions that are difficult to recover from.

Perhaps the most important result in Figure 4, however, is the large overlap between the participant and random distributions. One possible reason for this is that many participants may be making decisions that can be described by simple heuristic strategies, and are not necessarily adhering to the optimal decision process. Of course, it is not the only possibility, because a poorly executed optimal strategy could perform worse than a good simple heuristic that is executed accurately. Nevertheless, the overlap evident in Figure 4 needs attention. This is because, for inferences about the assumptions participants make about the environment—through the α and β parameters—to be useful, it is important to have evidence that their decisions are consistent with the assumed optimal decision process. For this reason, we decided to develop a modeling analysis of the decision-making data that allows for the evidence for adherence to the optimal decision process to be assessed.

Modeling Analysis

Evidence Via Model Comparison

One way to measure the evidence the decisions made by participants provide for following the optimal decision process is to propose alternative accounts of how they completed the task, and use model selection methods to infer which of these is the most likely from their behavioral data. To this end, we considered three heuristic models of how people might solve bandit problems, and contrasted them with the optimal Bayesian model. The goal in developing these models was to build a sequence starting from an extremely simple guessing model, and finishing at the optimal model, and including between models with an intermediate level of psychological sophistication. In this way, it is possible to assess the human decision-making we observed in terms of not only whether it is consistent with optimality, but also the nature of the shortcomings when it is less than optimal.

Our first model was an extremely simple *guessing* model, without any parameters, to act as a baseline or anchor. According to this model, every decision on every trial is made by choosing one of the alternatives at random, with equal probability given to each possible choice. This guessing model has likelihood function

$$p(D_k^g = i \mid M_{\text{guess}}) = \frac{1}{N}, \tag{5}$$

where $D_k^g = i$ means the ith alternative is chosen on the kth trial of the gth game.

Our second model is a version of the classic *win-stay lose-shift* model (Robbins, 1952). Under this heuristic, the decision-maker chooses randomly on the first trial, and subsequently stays with an alternative with (high) probably λ if they receive a reward, but switches randomly to another alternative with probability λ if they do not receive a reward. This model—essentially, but for the random choice on the first trial—has likelihood function

$$p(D_k^g = i \mid D_{k-1}^g = j, \lambda, M_{\text{wsls}}) = \begin{cases} \lambda & \text{if } j = i \text{ and } R_{k-1}^g = 1 \\ (1-\lambda)\,/\,(N-1) & \text{if } j \neq i \text{ and } R_{k-1}^g = 1 \\ 1-\lambda & \text{if } j = i \text{ and } R_{k-1}^g = 0 \\ \lambda/\,(N-1) & \text{if } j \neq i \text{ and } R_{k-1}^g = 0, \end{cases} \tag{6}$$

where $R_{k-1}^g = 1$ means the $(k-1)$th trial of the gth game resulted in a reward. Notice that this model is cognitively very simple. Unlike the guessing model, it responds to whether or not rewards are provided, but does so based on only the proceeding trial, and so does not invoke any sort of memory.

Accordingly, our third model is more complicated, and assumes the decision-maker does relies on memory of previous trials. We call it the *success ratio* model, because decisions are made on the ratio of successes to failures for each of the alternatives at each trial. Formally, the model considers the ratio $(s_x + 1)\,/\,(s_x + f_x + 2)$ for the xth alternative, where s_x and f_x are the number of successes (i.e., rewards) and failures, respectively, for all the previous trials of the current game. After evaluating this ratio over all of the alternative, the model chooses the alternative having the greatest ratio with (high) probability δ. If two or more of the alternatives have the same greatest ratio value, the model chooses randomly between them. This model has likelihood function

$$p(D_k^g = i \mid \delta, M_{\text{sr}}) = \begin{cases} \delta/N_{\max} & \text{if } i \in \arg\max_x\,(s_x+1)\,/\,(s_x+f_x+2) \\ (1-\delta)\,/\,(N_{\max}) & \text{otherwise.} \end{cases} \tag{7}$$

We compare all four of the models—the guessing, win-stay lose-shift, success ratio, and optimal models— using Bayes Factors (Kass & Raftery, 1995). This is a standard method of Bayesian model selection, which compares the average likelihood of a participant's decisions being generated under one model rather than the other. Using the guessing model as a low-end benchmark, this gave us three Bayes Factors, involving the following ratios of marginal densities:

$$\begin{aligned} \text{BF}_{\text{opt}} &= \frac{p\,(D \mid M_{\text{opt}})}{p\,(D \mid M_{\text{guess}})} \\ &= \frac{\int p\,(D \mid \alpha, \beta, w, M_{\text{opt}})\,p\,(\alpha,\beta)\,p\,(w)\ \mathrm{d}\alpha\ \mathrm{d}\beta\ \mathrm{d}w}{p\,(D \mid M_{\text{guess}})} \\ &= \frac{\iiint \left[\prod_{g=1}^{G}\prod_{k=1}^{K} p\big(D_k^g \mid S_k^g, \alpha, \beta, w, M_{\text{opt}}\big)\right] p\,(\alpha,\beta)\,p\,(w)\ \mathrm{d}\alpha\ \mathrm{d}\beta\ \mathrm{d}w}{\prod_{g=1}^{G}\prod_{k=1}^{K} p\big(D_k^g \mid M_{\text{guess}}\big)}, \end{aligned} \tag{8}$$

$$
\begin{aligned}
\mathrm{BF}_{\mathrm{wsls}} &= \frac{p\left(D \mid M_{\mathrm{wsls}}\right)}{p\left(D \mid M_{\mathrm{guess}}\right)} \\
&= \frac{\int p\left(D \mid \lambda, M_{\mathrm{wsls}}\right) p\left(\lambda\right)\ \mathrm{d}\lambda}{p\left(D \mid M_{\mathrm{guess}}\right)} \\
&= \frac{\int \left[\prod_{g=1}^{G} \prod_{k=1}^{K} p\big(D_k^g \mid R_{k-1}^g, \lambda\big)\right] p\left(\lambda\right)\ \mathrm{d}\lambda}{\prod_{g=1}^{G} \prod_{k=1}^{K} p\big(D_k^g \mid M_{\mathrm{guess}}\big)}, \qquad (9) \\
\mathrm{BF}_{\mathrm{sr}} &= \frac{p\left(D \mid M_{\mathrm{sr}}\right)}{p\left(D \mid M_{\mathrm{guess}}\right)} \\
&= \frac{\int p\left(D \mid \delta, M_{\mathrm{sr}}\right) p\left(\delta\right)\ \mathrm{d}\delta}{p\left(D \mid M_{\mathrm{guess}}\right)} \\
&= \frac{\int \left[\prod_{g=1}^{G} \prod_{k=1}^{K} p\big(D_k^g \mid \delta\big)\right] p\left(\delta\right)\ \mathrm{d}\delta}{\prod_{g=1}^{G} \prod_{k=1}^{K} p\big(D_k^g \mid M_{\mathrm{guess}}\big)}. \qquad (10)
\end{aligned}
$$

We estimated these three BF measures for every participant, using brute-force methods based on a finite grid to approximate the required integrals. This grid had 40 evenly-spaced points over the domain $(0, 1)$ for the w, λ and δ parameters, and both the α and β parameters were sampled from 0.2 to 5 in increments of 0.3.

Model Recovery Ability

To test the usefulness of the Bayes Factor model selection measures, and the general identifiability of our models, we conducted a simulation study, using the same set of problems encountered by our participants. We generated four artificial decision-making data sets by simulating 451 participants, with each of the sets corresponding to one of the guessing, win-stay lose-shift, success ratio, and optimal models.

Because we conceive of the w, λ and δ parameters as "accuracy of execution" parameters—describing how an essentially deterministic decision process is translated to noisy observed choice behavior—we set them to the value 1 in generating artificial data. This corresponds to ideal execution of the decision strategy.[4] For the optimal model, we also need to choose values for the assumed nature of the environment, and here we chose to set $\alpha = \beta = 2$.

For every simulated participant in every condition, we then made two analyses of their decision data. First, we calculated the $\log \mathrm{BF}$ measures, to find which of the four models was best supported. Here we used uniform priors over the w, λ and δ parameters. Then, we found the maximum a posteriori (MAP) estimates of the parameters of the best supported model, unless the best model was the parameter-free guessing model.

[4] In separate simulation studies not reported here, we also tested the ability to recover artificial data generated by the win-stay lose-shift and success ratio models with λ and δ values ranging uniformly between 0 and 1, and found parameter recovery to be excellent.

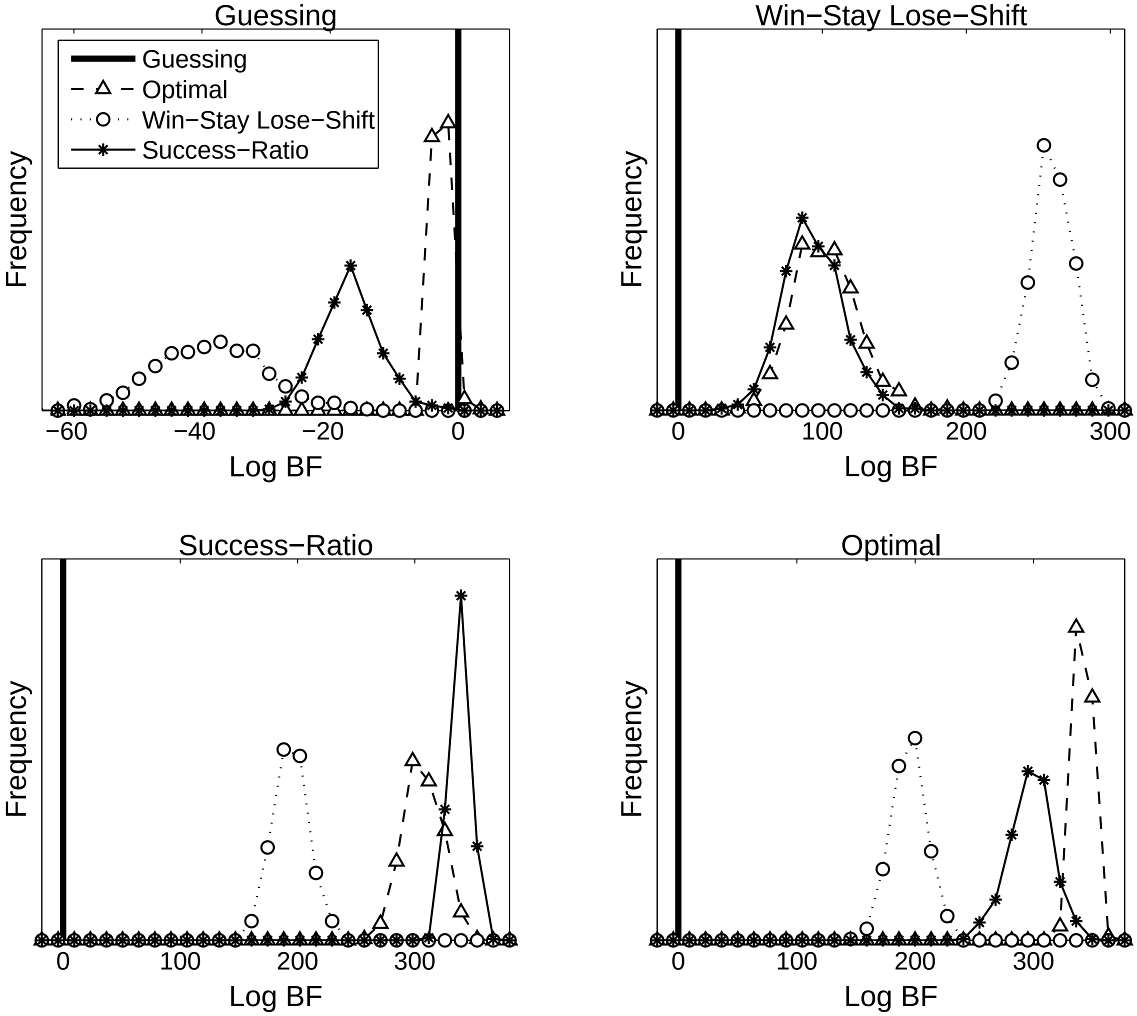

Figure 5. Results of the artificial model recovery analysis, showing the distribution of the three log BF measures over simulated participants whose data were generated by each of the four models.

Table 1: Model recovery performance, showing the proportion of simulated participants inferred to belong to each model for each of the four generating data sets.

Generating Model	Guessing	Win-Stay Lose-Shift	Success Ratio	Optimal
Guessing	98	0	0	2
Win-Stay Lose-Shift	0	100	0	0
Success Ratio	0	0	100	0
Optimal	0	0	0	100

Figure 5 summarizes the results of the first step in this analysis. Each of the four panels shows the distribution over all participants of the three log BF measures, relative to the guessing model, which by definition always sits at zero. A positive log BF value for one of the other three models constitutes evidence that participants were using that model rather than the guessing model, and the greater the values between competing models, the better supported they are.

It is clear from this analysis that our inference methods are able to recover the underlying decision models well, since there is little overlap between the best distribution and its competitors in each case. The exact level of recovery cannot, however, be gleaned from Figure 5 alone, because information about where each individual simulated participant lies within the three distributions is not shown. Accordingly, Table 1 presents the participant-specific details of the recovery. Over all of the simulations, the correct model *always* had the most evidence for data generated using the optimal, win-stay lose-shift and success ratio models. The guessing model data was correctly recovered 98% of the time, with the remaining 2% being attributed to the optimal model, as can be seen in the top-left panel of Figure 5. These results means we can have confidence applying the log BF measure to participants in our real behavioral data.

Figure 6 summarizes the results of the second step in the model recovery analysis, showing the distribution of MAP parameter estimates over simulated subjects, *conditional* on the model being the one recovered by the log BF measures. The large panel shows the distribution of α and β parameters of the optimal model in terms of the psychologically interpretable optimism $\alpha/\left(\alpha+\beta\right)$ and confidence $\alpha+\beta$ parameterization. The true environmental values $\alpha^{*}=\beta^{*}=2$ in this parameterization are shown the cross. The circles indicate the distribution of participants MAP estimates, with the area of the circle being proportion to the number of participants at each possible point in the parameter space. Recovery is clearly very good, although there is some minor deviation from the true generating values for a few simulated subjects. The smaller panels in Figure 6 show the distribution of the w, λ and δ accuracy of execution parameters. It is clear that these are perfectly recovered.

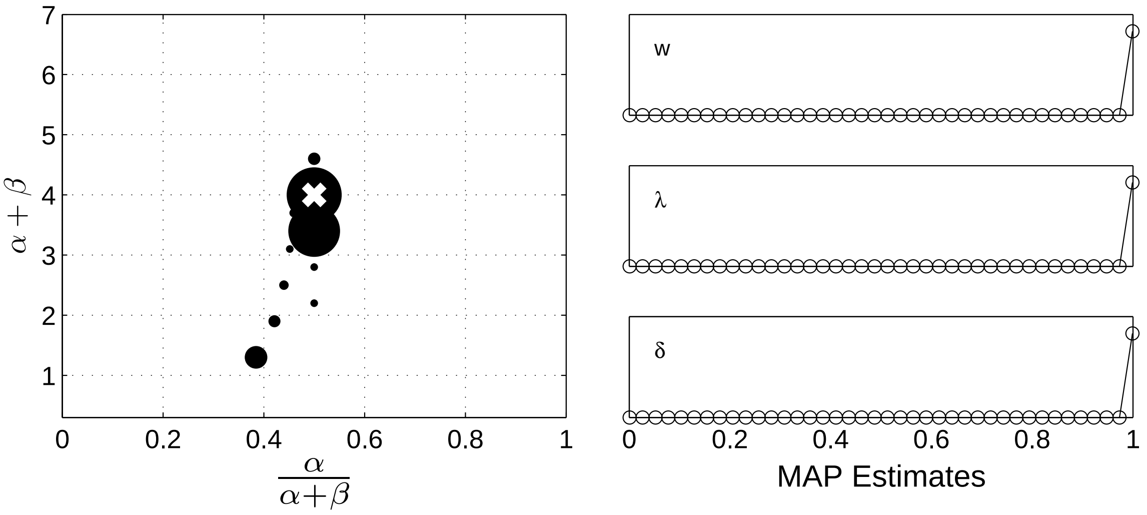

Figure 6. Distributions of parameter estimates in the artificial model recovery analysis. The large panel relates to the α and β parameters of the optimal model, with the true generating value shown by the cross. The three smaller panels relate to the w, λ and δ accuracy of execution parameters.

Modeling Results

In this section, we apply the models to the human data, in order to make inference about which model and parameterization best describes their decision-making. We then use information from these inferences to explore the relationship between individual decision-making and various psychometric and personality variables.

Bandit Problem Decision-Making

We begin by applying the log BF measures to the data for each of the 451 human participants, assuming uniform priors on the w, λ and δ parameters. The results are summarized in Figure 7. The left panel shows the distribution of log BF measures. The right panel shows the break-down of the participants into the proportions who best supported each of the four models (i.e., the MAP model estimate, assuming equal priors).

As anticipated from our original analysis of average reward distributions, there is clear evidence of individual differences in Figure 7, with a significant proportion of participants being most consistent with all three of the optimal, success ratio, and win-stay lose-shift models. Interestingly, about half of our participants were most consistent with the psychologically simple win-stay lose-shift strategy, while the remainder were fairly evenly divided between the more sophisticated success ratio and optimal models. Very few participants provided evidence for the guessing model, consistent with these participants being 'contaminants', who did not try to do the

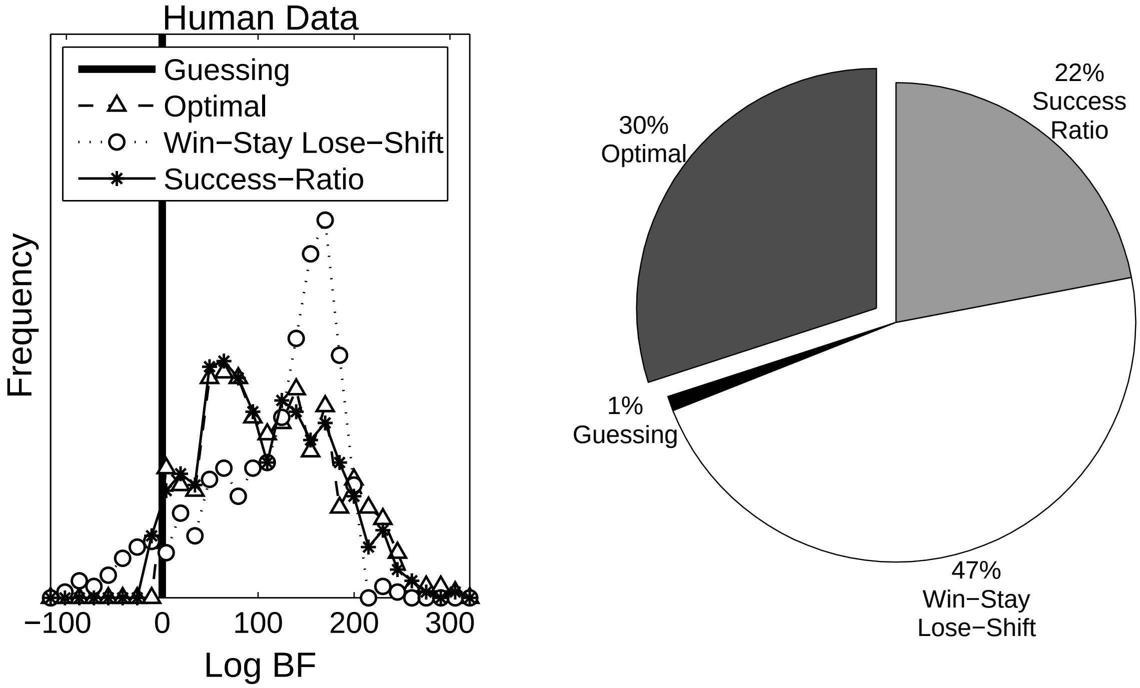

Figure 7. The distribution of the log BF measures over the participant data (left panel), and the sub-division into the proportions best supported by each of the four models (right panel).

task.

One interpretation of these results is that subsets of participants use successively more sophisticated decision-making strategies. The win-stay lose-shift decision model does not involve any form of memory, but simple reacts to the presence or absence of reward on the previous trial. The success ratio model involves comparing the entire reward history of each alternative over the course of the game, and so does require memory, but is not explicitly sensitive to the finite horizon of the bandit problem. The optimal model is sensitive to the finite horizon, and to the entire reward history, and so involves trading off exploration and exploitation. Figure 7 suggests that sizeable subsets of participants fell at each of these three levels of psychological sophistication.

Figure 8 shows the distribution of MAP parameter estimates, *conditional* on the model with the most Bayes Factor evidence. The three smaller panels in Figure 8 show that, with the occasional exception of the w parameter for the optimal model, the inferred accuracy of execution was generally high. More interestingly, the large panel shows the distribution for the α and β parameters of the optimal model, in the $\alpha/(\alpha+\beta)$ and $\alpha+\beta$ parameterization. It is clear that the participants behaving consistently with the optimal model show evidence of large individual differences. A

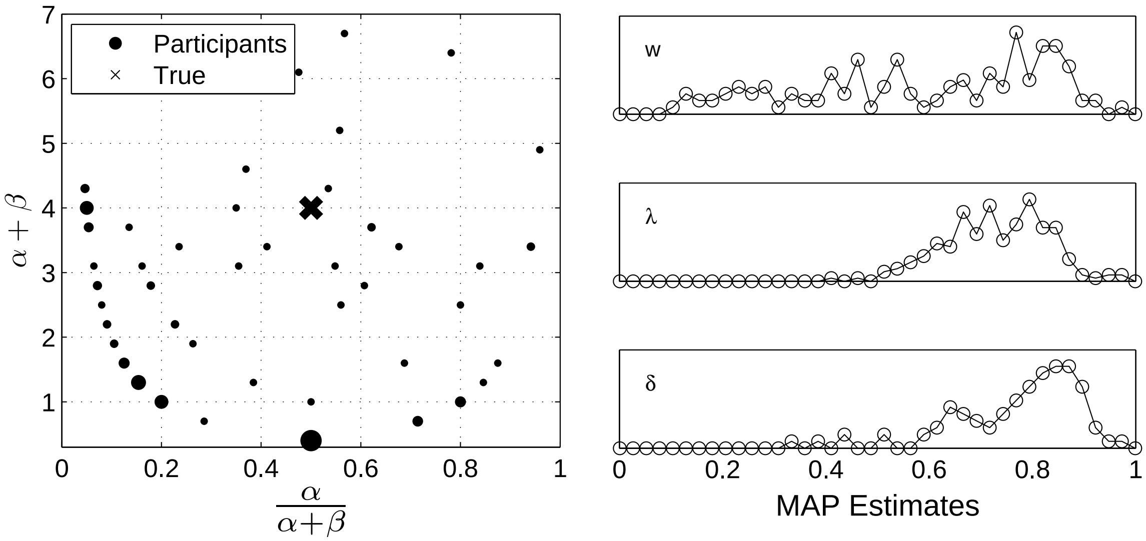

Figure 8. MAP estimates of the parameters for the optimal, success-ratio and win-stay-lose-shift models, based on just those subjects whose behavior was best accounted for by each of these models.

broad range of average reward rates $\alpha/(\alpha+\beta)$ is spanned by the MAP estimates. The certainty $\alpha+\beta$ estimates are generally low (and lower than the true state of the environment), but some participants behave consistently with assuming much larger values.

A sensible question to ask how the key result in Figure 7—which shows how participants are allocated to the four models—might be affected by different prior assumptions about the responding parameters w, λ and δ. These parameters are conceived as measuring the probability that the optimal, win-stay lose-shift, and success-ratio deterministic models will be followed on any trial. While we do not expect these accuracies of executions to be perfect, we do expect them to be relatively high. In this sense, placing a uniform prior on the w, λ and δ does not match our expectations well.

To address this issue, we repeated the $\log \mathrm{BF}$ analysis for eight other prior distributions. These were the $\mathrm{Beta}(2,2)$, $\mathrm{Beta}(4,4)$, $\mathrm{Beta}(2,1)$, $\mathrm{Beta}(4,2)$, $\mathrm{Beta}(8,4)$, $\mathrm{Beta}(4,1)$, $\mathrm{Beta}(8,2)$ and $\mathrm{Beta}(16,4)$ distributions. Using the interpretation of the parameters of the Beta distribution as 'prior successes' and 'prior failures', it is clear the last six variants give more weight to accurate over inaccurate execution of the models. As a set, these prior distributions correspond much more closely to a reasonable range of assumptions that might be made about the accuracy of execution parameters.

The results of the $\log \mathrm{BF}$ analysis with the alternative priors, together with the original analysis using the uniform $\mathrm{Beta}(1,1)$ prior, are shown in Figure 9. Each

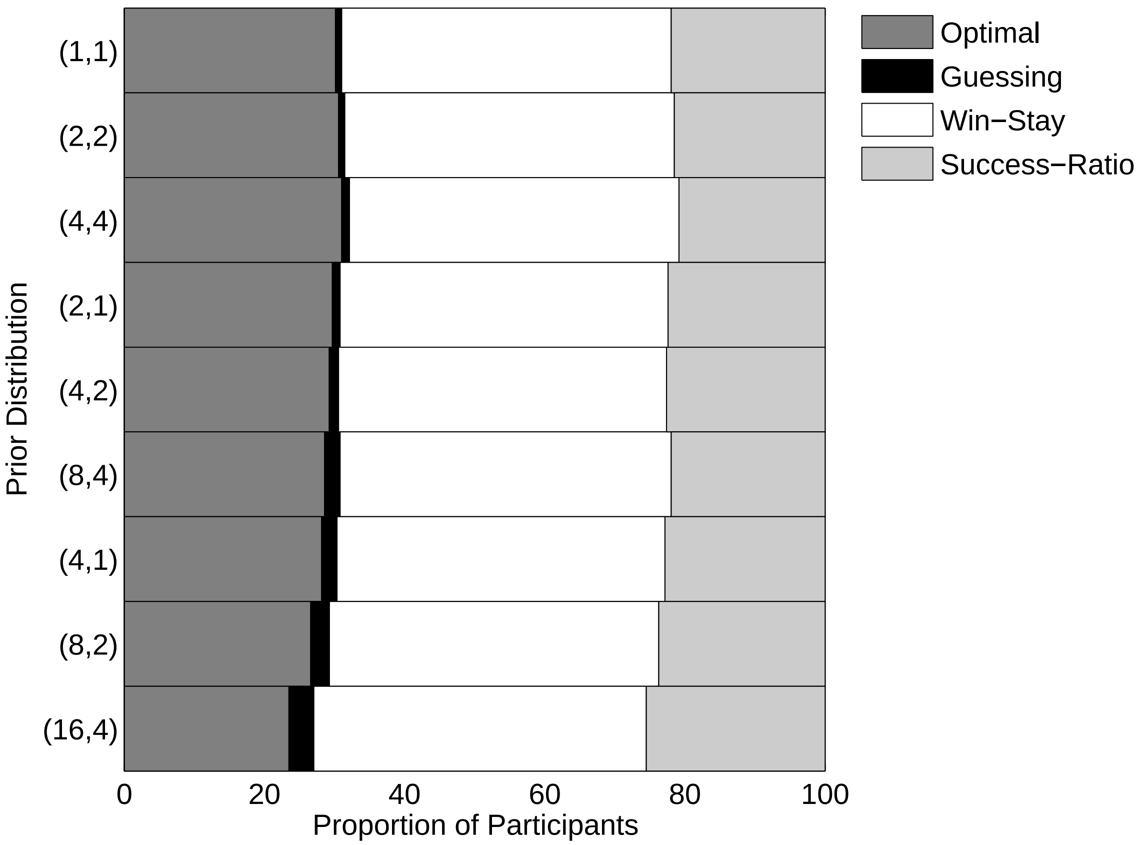

Figure 9. Proportion of participants best supported by each of the four models, using the log BF measure with different prior assumptions about the accuracy of model execution.

stacked bar graph running horizontally corresponds to a prior distribution, and the areas for each model correspond to the number of participants inferred to belong to that model. It is clear that the different prior assumptions about the accuracy with which people apply the models makes only minor differences to the assignment proportions. The basic conclusion that there are many win-stay lose-shift participants, some optimal and success-ratio participants, and few guessing participants, holds over the entire range of plausible prior assumptions.

Relationship to Psychological Variables

To examine the relationship between how all of the participants made decisions on the bandit problem, and other psychological traits measured as part of the test battery, we considered a set of 16 variables that characterize various interesting aspects of how the bandit problems were solved. These variables are defined in Table 2. They include a series of variables—explore, exploit, better, worse and untried—that count the number of decisions made according to simple rules involving the state of the game.

The 'explore' count is the number of times, over all decisions in all games, that a participant chose an alternative with fewer successes and fewer failures than another alternative. For the sample game state shown in Figure 3, choosing the fourth alternative would correspond to an 'explore' decision in relation to alternatives two and three, because it has fewer successes and fewer failures than both. The 'exploit' count is the number of times a participant chose an alternative with more successes *and* more failures than another alternative. In Figure 3, choosing the second or third alternative would correspond to an 'exploit' decision in relation to the fourth alternative, because they both have more successes and more failures. The 'better' count is the number of times a participant chose an alternative with more successes *and* fewer failures than another alternative. In Figure 3, choosing the second, third, or fourth alternatives would correspond to a 'better' decision in relation to the first alternative, because they have more successes and fewer failures. The 'worse' count is the number of times a participant chose an alternative with fewer successes *and* more failures than another alternative. In Figure 3, choosing the first alternative would correspond to a 'worse' decision in relation to the second, third, or fourth alternatives, because it has more successes and fewer failures than the others. Finally, the 'untried' count is the number of times a participant chose an alternative that had never been chosen before. In Figure 3, this corresponds to choosing the fourth alternative.

The other decision variables in Table 2 include the average reward per trial achieved by the participant, which corresponds to the outcomes of the decisions, as seen by the participant, and the expected reward per trial for the decisions made by the participant, which corresponds to the underlying rate of reward for each alternative chosen, and so is not directly observed by the participant. The absolute difference between the expected reward in the environment, and that inferred from the MAP estimates is included as variable $d = |\alpha/\left(\alpha+\beta\right) - 0.5|$. The MAP estimate measures of optimism $\alpha/\left(\alpha+\beta\right)$ and certainty $\alpha+\beta$ are included, together with the MAP estimates of the w, λ and δ accuracy of execution parameters. Finally, the three $\log \mathrm{BF}$ measures that compare the optimal, success ratio and win-stay lose-shift models to guessing model are included.

We examined the relationship of each of the decision variables to a set of 12 psychological variables. These included basic demographic information such as gender and age; psychometric measures of intelligence, including a measure of general intelligence g and two versions of the Ravens progressive matrices (Raven, Court, & Raven, 1988); and measures of personality traits, including creativity, self control, change, extraversion, neuroticism, and rigidity, that seemed plausibly related to the risk attitude balance inherent in balancing exploration and exploitation.

Table 3 gives the correlations between each decision-making variables and each psychological variable. Once again, those correlations that are significantly different from zero using the bootstrapped 95% confidence interval are highlighted in bold.

Variable	Explanation
Explore	The number of times a participant chose an alternative with fewer successes *and* fewer failures than another alternative.
Exploit	The number of times a participant chose an alternative with more successes *and* more failures than another alternative.
Better	The number of times a participant chose an alternative with more successes *and* fewer failures than another alternative.
Worse	The number of times a participant chose an alternative with fewer successes *and* more failures than another alternative
Untried	The number of times a participant chose an alternative that had never been chosen before.
Return	The average reward per trial achieved by the participant.
Expected	The expected reward per trial for the decisions made by the participant.
Difference, d	The absolute difference between the expected reward in the environment, and inferred MAP estimates of α and β.
Optimism, $\alpha/(\alpha+\beta)$	The value of the expected reward in the environment inferred from the MAP estimates of α and β.
Certainty, $\alpha+\beta$	The certainty in the expected reward in the environment inferred from the MAP estimates of α and β.
Execution, w	The MAP estimate of the w accuracy of execution in the optimal decision model.
Execution, λ	The MAP estimate of the λ accuracy of execution in the win-stay lose-shift model.
Execution, δ	The MAP estimate of the δ accuracy of execution in the success ratio model.
$\log \mathrm{BF}_{\mathrm{opt}}$	The log Bayes factor measure comparing the optimal and guessing models.
$\log \mathrm{BF}_{\mathrm{wsls}}$	The log Bayes factor measure comparing the win-stay lose-shift and guessing models.
$\log \mathrm{BF}_{\mathrm{sr}}$	The log Bayes factor measure comparing the success ratio and guessing models.

Table 2: Explanations of the decision-making variables.

Table 3: Pearson product-moment correlations between each decision variables and each psychological variable. Bold entries indicate correlations with bootstrapped 95% confidence intervals that do not include 0. (Note: Gender was encoded as male=1, female=2).

	Explore	Exploit	Better	Worse	Untried	Return	Expected	d	$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	w	λ	δ	$\log \mathrm{BF}_{\mathrm{opt}}$	$\log \mathrm{BF}_{\mathrm{wsls}}$	$\log \mathrm{BF}_{\mathrm{sr}}$
Gender	**0.12**	**-0.12**	0.07	0.08	0.01	-0.04	0.01	0.06	0.01	**-0.12**	-0.08	0.06	**-0.11**	-0.09	0.05	**-0.11**
Age	-0.05	0.01	-0.01	0	-0.03	0.01	0.02	0.05	-0.02	-0.01	0.05	0.05	0.02	0.05	0.09	0.03
g	**-0.09**	0.03	0.05	-0.09	0.06	0.04	0.04	**-0.18**	**0.10**	0.06	**0.11**	0.07	**0.15**	**0.13**	-0.02	**0.13**
Ravens 1	-0.09	0.02	0.05	**-0.10**	0.05	0.04	0.03	**-0.14**	**0.09**	-0.01	0.08	0.06	0.08	0.08	-0.02	0.06
Ravens 2	-0.07	0.06	0.03	**-0.16**	0.02	0.05	0.03	**-0.11**	**0.09**	0.03	**0.16**	**0.13**	**0.16**	**0.15**	0.06	**0.13**
Creativity	0.02	0.01	-0.01	-0.01	0.05	0.03	0.01	-0.09	0.03	0	0.03	0.03	0.09	0.03	0.01	0.07
Self Control	-0.02	-0.02	0.07	-0.09	0.01	0.01	-0.02	0	0.03	-0.01	0.05	0.08	0.05	0.02	0.03	0.02
Change	0.08	-0.06	0	0.09	0.03	-0.03	-0.02	-0.03	0.03	-0.02	-0.06	-0.01	-0.06	-0.06	0	-0.06
Order	0.09	-0.03	0.01	-0.03	-0.01	-0.01	-0.02	0.01	-0.06	-0.06	0.02	0.06	0.03	0.02	0.07	0.02
Extraversion	0.03	-0.04	0.01	0.05	0.04	0.03	0.05	-0.06	0.03	-0.02	-0.04	0	-0.03	-0.02	-0.02	-0.02
Neuroticism	-0.01	0.07	**-0.11**	0.02	**-0.13**	-0.04	-0.01	**0.10**	-0.09	0.05	0.02	-0.02	-0.03	0.01	0.06	0.01
Rigidity	**0.11**	-0.08	0.00	**0.11**	0.03	**-0.10**	-0.04	0.04	0.01	**-0.11**	-0.09	0.01	**-0.11**	**-0.11**	0.03	**-0.12**

There are relatively few significant correlations, and those that are significant are not large. Accordingly, we view this analysis as exploratory and suggestive at best, and interpret the results with caution.

The most interesting finding is that the $\log \mathrm{BF}_{\mathrm{opt}}$ and $\log \mathrm{BF}_{\mathrm{sr}}$ measures have significant positive correlations with the g and Ravens 2 intelligence measures. This suggests the ability of participants to follow the optimal or a near-optimal decision process provides an indication of their general intelligence. We note that basic experimental measures, such as the achieved rate of return and expected rate of return, do not correlate with these intelligence measures, nor does evidence of using the very win-stay lose-shift decision strategy. This underscores the usefulness of the model-based approach we have used to characterize the decision-making abilities of the participants.

Table 3 also shows that participants with a higher intelligence—as measured by g and the two Ravens measures—show smaller deviations d from the true average reward rate. This means that more 'intelligent' participants are not only better able to approximate the optimal decision strategy, but they also better approximate to the true environmental parameters used in the experiment. The demographic variables show few interesting correlations.

Finally, we note that the relationships between bandit problem decision-making and personality variables are also surprisingly weak. One could have speculated, a priori, that participants associated with higher risk-seeking behavior, creativity, or neuroticism would show a preference towards exploration, but there are few discernable pattern in the correlations. The one possibility relates to rigidity, which is negatively correlated with the $\log \mathrm{BF}_{\mathrm{opt}}$ and $\log \mathrm{BF}_{\mathrm{sr}}$, and suggests that too much rigidity somehow prevents the adoption of optimal or near-optimal decision-making.

Discussion

The bandit problem provides an interesting decision-making task that is amenable to formal analysis, but realistic enough to engage many important issues in understanding how people search for information and make choices. In this paper, we have focused on the tradeoff between exploration and exploitation in decision-making that lies at the heart of the problem. In particular, we were interested in whether there are individual differences in how people make the tradeoff, and how any differences might be related to broader psychological constructs including intelligence and personality.

We developed an extension of the optimal decision process that allows for people to have different assumptions—ranging from optimistic to pessimistic—about the statistical structure of the reward rates of alternatives. Using Bayesian methods, we showed how it is possible to make inferences about people's environmental assumptions from behavioral data. Our artificial recovery studies showed that these methods work well, even with a relatively small number of data.

We think our results highlight the benefits of developing formal models to test empirical hypotheses. Models in the cognitive science can be thought of as mechanisms for related raw behavioral data to a (usually much smaller) set of latent and psychologically meaningful parameters that generated the behavior. Having the freedom to express alternative accounts of decision-making as statistical models, and using them to infer these underlying parameters, is a powerful combination. For example, using the Bayes Factor to compare behavioral evidence for optimal decision-making versus alternative heuristic models of simpler decision-making, allowed us to process a large but very variable data set in a principled way.

Even more compellingly, it is the model-based measures, and not the simple experimental measures, that provided the most insight in exploring the relationship people's bandit problem performance to their intelligence and personality measures. We found that standard psychometric measures of intelligence had significant correlations with a Bayes factor model selection measure of the optimality of people's decision-making, but this relationship was not detected by simpler measures like the achieved average return.

One important way to extend the work we report here is to consider a richer and larger set of possible models as accounts of people's decision-making. There are obviously many possibilities, but we want to highlight one we suspect is particularly important. This is the 'soft-max' extension of the success-ratio model we used, which introduces an additional parameter to control the exploration-exploitation tradeoff. This model is well established in the reinforcement learning literature (Kaelbling, 1993), and is still widely used (e.g, Daw et al., 2006). We expect that fitting this model to behavioral data would provide a parameter-estimation approach to understanding how people make decisions on bandit problems, and would complement our approach based applying model selection methods to a set of simple deterministic heuristics.

The other important way to extend the study of exploration and exploitation in bandit problems, building on the current Bayesian optimal model, is to consider learning. In this paper, we have assumed that people have one fixed assumption about the distribution of rewards in the environment. A more realistic assumption would be that, at least over the course of completing a large number of problems, people are able to learn from the task what their assumptions should be. If an environment proves to be plentiful, with many high reward rates, there should be a shift towards exploration. In scarce environments, there should be a shift towards exploitation. Our Bayesian model can be extended to have the capability to learn, in optimal ways, from feedback as decisions are made. In this way, it is possible to continue using a model-based approach to examine how people adjust exploration and exploitation as they are confronted with different sorts of environments.

Acknowledgments

This work was funded by award FA9550-07-1-0082 from the Air Force Office of Scientific Research. We are very grateful to two anonymous reviewers, and Jerry

Busemeyer, for their extremely helpful comments on an earlier version of this paper.

References

Anderson, C. M. (2001). *Behavioral models of strategies in multi-armed bandit problems.* Unpublished doctoral dissertation, California Institute of Technology.

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory*, *10*, 55–77.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7–15.

Berry, D. A. (1972). A Bernoulli two-armed bandit. *The Annals of Mathematical Statistics*, *43*(3), 871–897.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments.* London: Chapman & Hall.

Brand, H., Sakoda, J. M., & Woods, P. J. (1957). Effects of a random versus pattern reinforcement instructional set in a contingent partial reinforcement situation. *Psychological Reports*, *3*, 473–479.

Brand, H., Wood, P. J., & Sakoda, J. M. (1956). Anticipation of reward as a function of partial reinforcement. *Journal of Experimental Psychology*, *52*(1), 18–22.

Brezzi, M., & Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics & Control*, *27*, 87–108.

Burns, N. R., Lee, M. D., & Vickers, D. (2006). Individual differences in problem solving and intelligence. *Journal of Problem Solving*, *1*(1), 20–32.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*, 253–262.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 933–942.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (Second ed.). London: Chapman and Hall.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal*, *61*, 35–73.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, *41*, 148–177.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices.* New York: Wiley.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Grünwald, P. D. (1998). *The Minimum Description Length Principle and Reasoning UnderUncertainty.* University of Amsterdam: Institute for Logic, Language and Computation.

Horowitz, A. D. (1973). *Experimental study of the two-armed bandit problem.* Unpublished doctoral dissertation, The University of North Carolina, Chapel Hill, NC.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kaelbling, L. P. (1993). *Learning in embedded systems.* Cambridge, MA: MIT Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitivemodels. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 807–812). Mahwah, NJ: Erlbaum.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555-580.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York, NY: Wiley.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on evolutionary computation*, *2*(1), 2-22.

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambuigity: Intuitive solutions to the armed-bandit problem. *Management Science*, *41*(5), 817–834.

Raven, J. C., Court, J. H., & Raven, J. (1988). *Manual for Raven's progressive matrices and vocabulary scales. section4: Advanced progressive matrices.* London: Lewis.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *55*, 527–535.

Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk–taking task. *Psychological Review*, *112*, 862-880.

Wood, S., Busemeyer, J., Koling, A., Cox, C. R., & Davis, H. (2005). Older adults as adaptive decision makers: Evidence from the Iowa gambling task. *Psychology and Aging*, *20*, 220–225.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision–making deficits. *Psychological Science*, *16*, 973–978.