M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 9, No. 4, Fall 2007, pp. 383–408
ISSN 1523-4614 | EISSN 1526-5498 | 07 | 0904 | 0383

informs®
DOI 10.1287/msom.1060.0130
© 2007 INFORMS

Simple Models of Discrete Choice and Their Performance in Bandit Experiments

Noah Gans
OPIM Department, The Wharton School, University of Pennsylvania, 3730 Walnut Street, Suite 500,
Philadelphia, Pennsylvania 19104, gans@wharton.upenn.edu

George Knox
Department of Marketing, Tilburg University, Room K 1019, P.O. Box 90153,
5000 LE Tilburg, The Netherlands, g.knox@uvt.nl

Rachel Croson
OPIM Department, The Wharton School, University of Pennsylvania, 3730 Walnut Street, Suite 500,
Philadelphia, Pennsylvania 19104, crosonr@wharton.upenn.edu

Recent operations management papers model customers as solving multiarmed bandit problems, positing that consumers use a particular heuristic when choosing among suppliers. These papers then analyze the resulting competition among suppliers and mathematically characterize the equilibrium actions. There remains a question, however, as to whether the original customer models on which the analyses are built are reasonable representations of actual consumer choice. In this paper, we empirically investigate how well these choice rules match actual performance as people solve two-armed Bernoulli bandit problems. We find that some of the most analytically tractable models perform best in tests of model fit. We also find that the expected number of consecutive trials of a given supplier is increasing in its expected quality level, with increasing differences, a result consistent with the models' predictions as well as with loyalty effects described in the popular management literature.

Key words: service quality; multiarmed bandit; decision-making under uncertainty; discrete-choice models; experimental tests
History: Received: December 16, 2004; accepted: August 30, 2006.

1. Introduction

In many business contexts, customers switch among suppliers. This switching is often observed in service industries, where it is called customer "defection" or "churn." It is also a common phenomenon in the context of consumer products, where brand switching is a widely studied phenomenon. Among the factors responsible for this switching is random variation in the utility or value that a customer obtains each time he or she patronizes a supplier.

More specifically, in many settings, the quality of service offered to customers has an inherently random component. Competing suppliers' quality distributions, in turn, jointly determine customer switching behavior and market shares. By installing extra capacity or instituting additional quality-control measures, a service provider can improve—at a cost—the distribution of its service quality.

The mechanism by which customer switching occurs is naturally modeled as the solution to a multiarmed bandit problem. While the bandit problem is a straightforward means of representing a customer's repeated choices, it is difficult analytically to work with, and this difficulty raises two sets of questions. Customers faced with bandit problems need practical characterizations of optimal (or near-optimal) behavior so that they make better decisions. Companies—the "arms" in consumers' bandit problems—need to better understand how customers actually make choices in a bandit setting so companies can effectively respond to their needs.

A number of recent operations management papers have recommended actions to companies that are arms in bandit or bandit-like problems (e.g., Hall and Porteus 2000, Gans 2002a, Gaur and Park 2003). These papers begin by positing that customers use a particular heuristic for making repeated choices under uncertainty. They then derive aggregate statistics regarding customers' choices, such as each supplier's "customer share," as a function of a given set of supplier quality distributions.[1] Finally, these papers model suppliers as competitively choosing quality distributions, and they characterize suppliers' resulting equilibrium choices.

A common feature of these papers is the analytical tractability of their choice models. That is, the rules customers are assumed to follow are not complex, and this allows the resulting expressions for aggregate choice statistics to be simple functions of firms' quality distributions.[2] While this simplicity facilitates the papers' competitive analyses, there remains a question as to whether or not the underlying customer models are reasonable representations of actual consumer choice.

In this paper, we report one effort's results at validating these models of consumer choice. We view the models we test as representing a variety of tradeoffs between analytical tractability and richness of the representation of customers' learning processes. To the extent that more stylized, tractable models adequately capture customer choice behavior, they may be valuably used in the type of competitive analysis described above. Alternatively, if more complex models—whose statistics are more difficult to derive—are significantly better predictors of customer behavior, then the use of highly stylized models needs to be rethought.

We consider two sets of models. One set includes three models that represent successive approximations of a normative analysis of the problem. The first is a Gittins index model, which posits that subjects make choices to maximize the expected discounted value of their outcomes and use Bayes's rule to incorporate learning from past experience. The next is a Myopic analogue to the Gittins index rule, in which subjects are Bayesian but choose to maximize the expected value of only their immediate choice. The last model, which we call Simple, further reduces the model of myopic behavior by assuming that subjects categorize arms as being either "good" or "bad."

The second set includes three related models of choice that are outgrowths of the literature on statistical learning and decision making under uncertainly. The first is a version of the myopic rule, described above, in which subjects remember the outcome of only the last n trials. A further simplification is a hot hand (HH) rule, in which subjects stay with the current arm until it loses in n consecutive trials. The last is an exponential smoothing (ES) model, in which subjects update their beliefs about the quality of an arm by taking a weighted average of their past beliefs with the current outcome. In §2, we review the literature related to the proposed models and their tests, and in §3, we formally define the two sets of models to be analyzed.

We use experiments to test the performance of the models, and we analyze the results in two ways. First, we consider a prediction consistent across the representations: The expected switching time (expected number of trials that a subject consecutively samples from a given arm before switching to an alternative) should be increasing and convex in the expected payout of the arm. The increasing-convex property is also consistent with the claim, made in the popular management literature, that marginal increases in service quality can have increasingly dramatic benefits in customer loyalty and lifetime value (Jones and Sasser 1995).

Second, we distinguish among the various models' ability to match subjects' behavior, choice by choice. The use of experiments to distinguish among models in a horse race like this is common (Roth 1988). The experimental setting must be carefully constructed so that all models make clear predictions and a dimension exists along which they can be evaluated. Our experiments, described in §4, are designed to do just this.

Our findings, presented in §§5 and 6, are as follows. First, the expected switching time behaves as predicted as a function of the expected payout of an arm;

[1] Customer share refers to the fraction of all purchases that a customer makes at a given supplier.

[2] It is worth noting that a model's analytical tractability can be viewed in at least two ways. On the one hand, one may view as more tractable a model that requires less computation or memory on the part of subjects. On the other hand, one may view a model as more tractable if aggregate measures of a subject's performance are simple functions of system parameters, preferably with simple, closed-form mathematical expressions. We are primarily concerned with the latter, and when we write "tractable," this is what we mean. Nevertheless, typically, the two meanings go hand in hand.

the behavior is consistent with the increasing and convex properties implied in the literature. Second, more analytically tractable models provide the best fit to subjects' observed choices. In contrast, the more complex Gittins index model—which maximizes expected discounted rewards—fares most poorly in tests of model fit.

Our experimental results also suggest that no single model dominates the others. That is, there exist significant subgroups of subjects for whom different choice models perform best. This finding suggests that suppliers' choices of service-quality distribution(s) should explicitly account for various market segments, each segment characterized by a different model of choice. These conclusions and other extensions are discussed in §§7 and 8.

2. Literature Review

Bandit-like models have been used in a wide variety of contexts. Classic examples include clinical drug trials and petroleum exploration (e.g., Gittins 1989, Banks et al. 1997, Anderson 2001), and in consumer product or service settings (e.g., Meyer and Shi 1995, Erdem and Keane 1996). For a more complete description of the multiarmed bandit problem, see Appendix A.[3]

Because of its importance in many different fields, normative aspects of the bandit problem have received considerable attention over the years, and we briefly describe previous work devoted to settings like ours—an infinite-horizon, discounted version of the problem. Perhaps most widely known is the work of Gittins and Jones (1974) who proved the following: From each arm's Markov chain and current state, an index may be calculated, and at any stage it is optimal to choose the arm with the highest index. Gittins (1979) went on to develop explicit expressions for this so-called Gittins index, though in the context of Bayesian bandits (with arms that are members of exponential families of distributions), the index has proven to be tremendously difficult to calculate. In turn, Chang and Lai (1987) developed approximations to the Gittins index for Bayesian bandits, which they proved to be asymptotically optimal as the discount rate approaches one, so that the decision maker samples infinitely before switching away from an arm (see also Brezzi and Lai 2002).

The difficulty in calculating optimal choices has also led to suggestions for simpler, heuristic solutions to the bandit problem. An early paper by Schmalensee (1975) analyzes the use of the Bush and Mosteller (1955) linear learning model for solving the problem. In the computer science literature, various other forms of reinforcement learning have been applied to the problem as well (Sutton and Barto 1998).

This difficulty naturally leads one to wonder how well people actually solve bandit problems. Surprisingly, however, relatively little positive work investigates how people make choices in bandit settings. In particular, we are aware of only four such papers. Horowitz (1973), Meyer and Shi (1995), and Banks et al. (1997) use Bernoulli bandits, whose arms have simple win-lose payout distributions. Anderson (2001) uses arms with more complex payout distributions, such as simulated dice rolls and normally distributed rewards.

Results from these studies suggest that people's behavior is roughly consistent with predictions from the Bayesian model, but with important and systematic deviations. Horowitz (1973) and Meyer and Shi (1995) offer exploratory analyses, using experimental data to generate hypotheses concerning the heuristics used by subjects. Both papers find that subjects' choices reflect a Bayesian updating of priors, but also a bias toward choices more myopic than are optimal. Horowitz (1973) also finds that subjects oversample from inferior arms (overexperiment), a bias that is distinct from that of myopic behavior. Banks et al. (1997) and Anderson (2001) devise experiments that provide sharp tests of specific hypotheses concerning the presence of certain heuristics or biases in choice behavior. Anderson (2001) finds that subjects experiment less than optimally and provides evidence that risk aversion associated with more diffuse priors is the likely cause of the effect. Banks et al. (1997) manipulate the arms' prior distributions so that, in some cases, optimal choice behavior is myopic and in others it is not, and they find that subjects' behavior is consistent with these normative predictions.

The intent of our analysis differs from the work described above. Rather than using the data either

[3] All appendices are available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

in an inductive fashion, to propose new models of choice, or to construct sharp tests of hypotheses concerning specific biases away from expected utility-maximizing behavior, we take an "engineer's" view of the problem. Our aim is to identify models of choice behavior that can easily render aggregate statistics useful in the context of competitive analysis (such as expected switching time or fraction of times chosen). Thus, we judge a model's value along two dimensions: its analytical tractability (for operations management professionals) as well as its ability to represent a wide variety of people's choices.[4]

Our approach is similar in spirit to recent work in the economics literature by Harless and Camerer (1994) and Camerer and Ho (1999). In the former, the authors consider generalizations of expected utility theory along two dimensions: "fit" with the observed data, as well as "parsimony" (parallel to our notion of tractability). In the latter, the authors apply tools from marketing research to judge the fit of models to subjects' choices in experiments, much as one would fit brand-choice models to panel data. We will pursue a similar strategy here.

The models we test come from a variety of sources. Two (Myopic and Simple) are simplifications of the original Gittins index model. These models are motivated by two sets of findings from empirical work on perception and decision making. The first observation is that people tend to choose more myopically than is optimal (e.g., Horowitz 1973, Meyer and Shi 1995). The Myopic model explicitly incorporates this bias. The second is that people appear to systematically categorize as they make sense of their experiences. That is, they maintain mental examples of how entities in the world behave, and they interpret an experience with an entity by comparing their perceptions with the typical or exemplary characteristics of their mental picture of how that category behaves. This structure appears in Kahneman and Tversky's well-known "representativeness heuristic" (1973, Tversky and Kahneman 1974), and it forms the basis of category and exemplar theory in cognitive psychology (e.g., Henderson and Peterson 1992). The Simple model assumes that subjects categorize arms as "good" or "bad" and that subjects exhibit the same myopia present in the Myopic model.

Models in the second set we tested are variants of those with origins in behavioral decision theory. Our "Last-*n*" model is based on the finding that bounds on memory limit subjects' ability to recall long histories of previous outcomes (Miller 1956). The "HH" model, described in Gilovich et al. (1985), posits that subjects erroneously ascribe positive serial correlation to the outcomes of repeated trials. (The mirror of this is the negative correlation implied in "gambler's fallacy" rules; e.g., Burns and Corpus 2004.) The classic, exponential smoothing model—first developed in the early 1960s (Brown and Meyer 1961)—was introduced into the marketing literature in Guadagni and Little (1983) and more recently proposed as a choice model in March (1996).

3. The Bandit Problem and Choice Models

In this section, we formally define the multiarmed bandit problem and the models of choice that we will test. For more detail on the bandit problem, see Appendix A (online).

3.1. The Multiarmed Bandit

The multiarmed bandit problem used in our study is defined as follows. There exist m arms, indexed $i = 1, \ldots, m$. When sampled, arm i yields a randomly distributed reward (or value or utility) with fixed distribution, U^i. Formally, we describe U^i as uniquely defined by a parameter $\theta \in \Theta$ and as having cumulative distribution function $F(u \mid \theta)$, so that $U^i \sim F(u \mid \theta^i)$. Let $\mu_\theta = \mathsf{E}[U \mid \theta] = \int u\, dF(u \mid \theta)$.

A subject must repeatedly choose among the m arms, and let $t = 1, 2, \ldots$ be the time index of his or her choices. While the subject knows that each arm has a fixed distribution $\theta^i \in \Theta$, he or she does not know what the various θ^i's are. The task is to repeatedly choose among the arms in a way that maximizes the aggregate value of choices.

Let a *policy* $\pi = \{\pi(1), \pi(2), \ldots\}$ be a sequence of choices of arms, and let Π be the class of policies

[4] Hutchinson and Meyer (1994) suggest that a formal positive theory of sequential choice is likely to include combinations of simple rules or strategies, each of which optimizes a limited or restricted version of the task at hand. In this context, one may view our analysis as explicitly considering the efficacy of various simple policies, while postponing the larger issue of how people use the rules in concert.

that is nonanticipating with respect to future rewards. Next, we formally define the subject's problem as that of finding (and executing) a policy, $\pi \in \Pi$, that will maximize the expected discounted value of the future stream of rewards, $\sup_{\pi\in\Pi} \mathsf{E}_\pi[\sum_{t=1}^{\infty} \alpha^t U_t]$, where $\alpha \in (0, 1)$ is the one-period discount rate.

3.2. The Gittins Index for the Bayesian Bandit and Related Models

A Bayesian subject may view each arm's θ^i as a random variable with support on Θ. For arm i, he or she maintains a cumulative distribution function, $P_t^i(\theta)$, that represents the understanding at time t of the quality distribution, $\theta^i \in \Theta$, under which he or she believes the arm to be operating. Let $\{P_0^1, \ldots, P_0^m\}$ be the "prior" information the subject has before he or she begins the sequence of choices.

After each choice, the Bayesian subject uses the new sample, U_t, and Bayes's rule to update his or her beliefs. Specifically, if he or she uses arm i at time t and receives reward u, then the new (posterior) belief distribution will be

$$dP_t^i(\theta \mid u) = \frac{dP_{t-1}^i(\theta) dF(u \mid \theta)}{\int_\theta dP_{t-1}^i(\theta)\, dF(u \mid \theta)} \quad \forall\, \theta \in \Theta. \tag{1}$$

If he or she does not sample from i, then $dP_t^i(\theta) = dP_{t-1}^i(\theta)\ \forall\, \theta \in \Theta$.

3.2.1. Gittins Index Model. The so-called Gittins index of arm j describes the expected discounted reward per unit of expected discounted time,

$$G(P_t^j) \triangleq \sup_{\mathcal{T}>t} \left\{ \frac{\mathsf{E}\big[\mathsf{E}\big[\sum_{s=t+1}^{\mathcal{T}-1} \alpha^{s-t} U(P_{s-1}^j) \mid P_t^j\big]\big]}{\mathsf{E}\big[\mathsf{E}\big[\sum_{s=t+1}^{\mathcal{T}-1} \alpha^{s-t} \mid P_t^j\big]\big]} \right\}, \tag{2}$$

where $\mathcal{T}$ is a stopping time with respect to the history of the process through time, $t-1$. The notation, $U(P_{s-1}^j)$, emphasizes that the marginal (subjective) distribution of value at $(s-1)$ is a function of the distribution of the subject's belief at the time. To maximize the expected discounted reward, a subject should choose the arm with the largest Gittins index, $i = \arg\max_j \{G(P_t^j)\}$ (Gittins and Jones 1974, Gittins 1979).

3.2.2. Myopic Model. In practice, $\mathcal{T}$ and the Gittins index are extremely difficult to compute, and experimental evidence suggests that people behave more myopically than is optimal (Horowitz 1973, Meyer and Shi 1995). In contrast, a *myopic* customer significantly simplifies the determination of the preferred arm by ignoring the option of future switching. For each arm j, he or she more simply calculates the expected discounted reward, given he or she remains on arm j for all time,

$$M(P_t^j) \triangleq \frac{\sum_{s=t+1}^{\infty} \alpha^{s-t} \mathsf{E}[U(P_{s-1}^j)]}{\sum_{s=t+1}^{\infty} \alpha^{s-t}} = \mathsf{E}[U(P_t^j)] = \int \mu_\theta \, dP_t^i(\theta). \tag{3}$$

He or she then chooses the arm i that maximizes this long-run expected reward: $i = \arg\max_j \{M(P_t^j)\}$. From the right side of (3), we see that this is equivalent to choosing the arm that myopically maximizes the expected reward at t.

Like the "rational" counterpart, the myopic subject uses the reward realized at i and Bayes's rule (1) to calculate her posterior beliefs, P_{t+1}^i, concerning i's quality distribution. Again, for all $j \neq i$, $P_{t+1}^j = P_t^j$.

3.2.3. Simple Model. While the myopic subject's task is significantly simpler than that of the Gittins index counterpart, it still may be quite complex. He or she must maintain a prior distribution over the set, Θ, and also perform potentially difficult integrations to update the prior distributions (1) and to calculate the indices (3).

Indeed, a common finding in behavioral research is that people may substitute simpler heuristics for these complex, integrative tasks. Perhaps the best-known example of this type of simplification is Kahneman and Tversky's "representative heuristic" (1973, Tversky and Kahneman 1974). The *Simple* model uses categorization to further simplify the integrative aspects of the myopic model.

In addition to using (3) to myopically choose the arm that maximizes immediate expected reward, a Simple subject partitions the arms' possible quality levels into two categories—good and bad—with respective reward distributions $F_G \equiv F(u \mid \theta^G)$ and $F_B \equiv F(u \mid \theta^B)$. That is, he or she further simplifies the choice process of the Myopic subject by reducing the set of distributions that he or she recognizes to be $\Theta = \{\theta^B, \theta^G\}$. In turn, P_t^i, the prior distribution he or she maintains for each arm's θ^i, collapses to be p_t^i, the probability that i is good rather than bad. Thus, rather than judging *how* good or bad an arm is, the subject's problem is more simply to decide *whether* an arm is good or bad.

Let $\mu_G \triangleq \mathsf{E}[U \mid \theta^G]$ and $\mu_B \triangleq \mathsf{E}[U \mid \theta^B]$. Using (3) we can define the index used by the Simple subject in terms that are exactly analogous to $M(P_t^i)$:

$$S(p_t^i) = \mathsf{E}[U(p_t^i)] = \mu_G p_t^i + \mu_B (1 - p_t^i). \tag{4}$$

Again, $S(p_t^i)$ is the expected reward of sampling from i at t. Then, given a realization, u, the Simple subject's use of Bayes's rule (1) to update the prior probability that i is good reduces to

$$\begin{aligned} p_t^i &= \frac{p_{t-1}^i dF_G(u)}{p_{t-1}^i dF_G(u) + (1 - p_{t-1}^i) dF_B(u)} \\ &= \left(1 + \frac{1 - p_{t-1}^i}{p_{t-1}^i} \cdot \frac{dF_B(u)}{dF_G(u)}\right)^{-1}. \end{aligned} \tag{5}$$

Algebraic manipulation shows that the index defined by (4)–(5) is equivalent to

$$\tilde{S}_{t-1}^i = X_0 + \sum_{s=1}^{t-1} 1\{\pi(s) = i\} \cdot X_s, \tag{6}$$

where $X_0 = \ln(p_0^i/(1 - p_0^i))$ is a log-likelihood that reflects the subject's initial belief concerning the probability that arm i is good (p_0^i), $X_s = \ln(dF_G(U_s)/dF_B(U_s))$ is the log-likelihood that the sth trial comes from a good arm, and $1\{\cdot\}$ is the indicator function. That is, $S(p_{t-1}^i) > S(p_{t-1}^j)$ if and only if $\tilde{S}_{t-1}^i > \tilde{S}_{t-1}^j$.

Note that (6) is a random walk that has an intuitive interpretation. In it, X_0 is an initial level of satisfaction that a subject has for arm i. Each time a subject samples from i, the experience leads an immediate response to the quality of the interaction, X_s. In turn, this immediate response is integrated into the subject's overall satisfaction with i in a straightforward, additive fashion: Better outcomes increase and worse outcomes decrease overall satisfaction.

Conversely, one can view the Simple model as being defined as (6), an additive model of "satisfaction" in which the arm with the highest cumulative satisfaction is chosen in each trial. This is precisely the model of "cumulative discrete choice" proposed in Gilboa and Pazgal (2001). Given this second definition, we then note that the Simple model is also consistent with the behavior of a myopic Bayesian who categorizes arms as good or bad.

The Simple model is also consistent with the behavior of a so-called "satisficing" subject. That is, rather than seeking the arm that maximizes expected rewards, the Simple subject will be satisfied with any arm that meets some target level of average reward per trial, μ^*. If $\mathsf{E}[U^i] < \mu^*$, then arm i does not meet the subject's required satisfaction level, and the expected number of times he or she will sample from i before switching is finite. If, however, $\mathsf{E}[U^i] \geq \mu^*$, then arm i meets the subject's so-called "aspiration level," and he or she is expected to continue sampling from i indefinitely. The twin notions of satisficing and aspiration levels have a long history, dating back to Simon (1959) and beyond.

The Simple model was used in Gans (2002a) in the context of models of service competition, and this paper provides a general, closed-form, representation of μ^* for exponential families of probability distributions. In §6, we also provide an explicit representation μ^* in the context of our experimental setting.

The models described above form a hierarchical family. Most complex is the Gittins index model, which derives from the rational behavior of Bayesian subjects. Less complex is the Myopic model, and then least complex is the Simple model, which may be viewed as a categorical version of the Myopic model, as well as an additive, "random walk" index of a subject's satisfaction with a given arm.

3.3. Other Choice Models

There exist many other possible representations of choice under uncertainty that can be tested in our bandit setting. In this paper, we concentrate on three: A "Last-n" model, which is analogous to a Myopic model with limited memory of past trials; an ES analogue to the Simple model; and an HH model that reacts to recent wins and losses on the currently sampled arm.

3.3.1. Last *n*. It is well known that individuals have only limited memory. For example, a long stream of research in psychology documents that individuals can remember roughly seven pieces of information, such as digits of telephone numbers, after which they need record-keeping or other external aids (Miller 1956).

The *Last-n* index explicitly incorporates the effect of limited memory, using only the results of the last n trials on a given arm. Formally, it is calculated in the same fashion as the Myopic index, the difference being that the Last-n index for an arm corresponds to a Myopic index in which only the previous n trials on that arm are remembered.

Although the model represents subjects as having limited memory of past events, in a significant sense, its use is more demanding of subjects' memory than the related Myopic rule. More specifically, let $k_i(t) = \sum_{s=1}^{t} 1\{\pi(s) = i\}$ be the number of times a subject has sampled arm i by time t, and let $s_i(k) = \min\{t \mid k_i(t) = k\}$ be the time of the kth sample from arm i. Then, a subject who uses the Last-n model must always recall each of the last n samples on each arm; that is, given he or she has sampled $j \geq n$ times from arm i, he or she must recall samples $\{U_{s_i(j-n+1)}, \ldots, U_{s_i(j)}\}$. In contrast, to update the analogous Myopic index, a subject need only recall the prior distribution, P^i_{t-1}, along with the outcome at time t, U_t.

3.3.2. Hot Hand. A subject that uses the HH model ascribes positive serial correlation to the trials of a repeated random sample. The rule's underlying premise is that an arm that has recently won is more likely (than average) to win again.[5]

We define an "HH-n" rule in our experiments as follows. If the subject experiences n consecutive losses in the last n trials on an arm, he or she should switch to the other arm; otherwise, he or she should continue to sample from the current arm. Thus, another way of stating an HH-1 rule is "stick on a winner and switch on a loser." Similarly, an HH-n rule could be described as switch only on n consecutive losers. Like the Last-n rule, the HH-n rule requires that a subject maintain a detailed record of the outcome of each of the last n trials on a given arm.

The HH-n family of rules differs significantly from the other rules that we test in two fundamental ways. First, as far as we can tell, it is only directly applicable to Bernoulli outcomes. While outcomes with more complex distributions can be reduced to Bernoulli trials by applying a threshold—so that outcomes above the threshold are considered to be wins and those below the threshold, losses—all the other rules immediately generalize to arbitrarily complex distributions, without having to be transformed in this manner. Second, the HH-n family is not an index rule. All the other rules that we test calculate an index for each arm and recommend that the subject sample from the arm with the higher index. In contrast, in the HH-n rule, the decision to stay on or switch away from the current arm is based only on that arm's performance. Information concerning the past performance of alternative arms is not used in the switching decision.

We note that Hall and Porteus (2000) use a variant of the HH-1 rule in their analysis of service competition. In this model, a customer who receives satisfactory service remains with the current supplier, while a customer who receives unsatisfactory service switches to a competitor with a fixed probability p.

3.3.3. Exponential Smoothing. Exponential smoothing is a weighted analogue of the Simple model's additive random walk. Formally, we define the ES model as follows. At $t = 0$ the subject's prior estimate of the average reward to be gained by sampling from arm i is $ES_i(0)$. Then, at each trial t, at which the subject samples from arm i, we let

$$ES_i(t) = \gamma U_t + (1-\gamma) ES_i(t-1)\,, \tag{7}$$

where $0 < \gamma < 1$. Again, for all $j \neq i$, $ES_j(t) = ES_j(t-1)$.

In contrast to the index for the Simple model, $\tilde{S}_i(t)$, for which each previous trial carries an equal weight, $ES_i(t)$ is more strongly affected by recent trials at i. To see this, again let $s_i(k)$ denote the time index of the kth trial at arm i. Then, given arm i has been pulled j times by time t, we can write $ES_i(t) = \sum_{k=0}^{j} \gamma(1-\gamma)^{j-k} U_{s_i(k)}$, where $U_{s_i(0)} \equiv \gamma^{-1} ES_i(0)$. Thus, for a fixed $0 < \gamma < 1$, the weight of the $s_i(k)$th trial at i declines geometrically (roughly exponentially) quickly with each new sample from i, hence, the name exponential smoothing. The higher the weighting factor, γ, the more quickly the weight declines. For $\gamma = 0$, the index remains constant, $ES_i(t) = ES_i(0)$ for all t. Conversely, for $\gamma = 1$, a subject's index is defined by his or her most recent trial: $ES_i(t) = \max_{\{s \leq t \mid \pi(s) = i\}} U_s$.

Exponential smoothing models have a long history of use in a number of fields related to learning. An early reference from the forecasting literature is

[5] The name "hot hand" derives from basketball, in which there is a common belief that players have "hot" hands, so that their success probabilities in making free-throw attempts exhibit positive serial correlation. In the context of problems of repeated choice under uncertainty, we say a belief in the "hot hand" implies that a subject (erroneously) believes that a future payout is (positively) serially correlated with recent performance (Gilovich et al. 1985).

Brown and Meyer (1961). Guadagni and Little (1983) is a well-known application from the marketing literature, which applies smoothing to a bandit-like brand choice problem. A recent example from the learning literature in psychology is March (1996), which uses simulation to analyze properties of smoothing models.

Of particular interest to us is a generalization of a smoothing model that appears in Gaur and Park (2003)—a paper that analyzes service-level competition among inventory systems. Here, outcomes are Bernoulli—either a customer order is filled or not—and the model is extended so that the smoothing constant, γ, may differ, depending on whether or not an order is filled. (This also echoes asymmetric elements of the stochastic learning models that date back to Bush and Mosteller 1955.)

It is also worth noting that the smoothing model is consistent with the belief that an arm's reward distribution is Markov modulated, rather than i.i.d. across trials. In this case, it can be shown that smoothing is analogous to the use of a Bayesian procedure in which prior trials' results are discounted (for details, see Matsuda and Sekiguchi 1971). In the context of our bandit problems, one may interpret a subject's use of exponential smoothing as "not believing" that arms are i.i.d. and, therefore, discounting earlier sample information, because it is more likely to have been obtained from a reward distribution that differs from the one currently being sampled.

3.4. Additional Models

There is a potentially vast array of additional models that one might also consider. For example, we have not analyzed models of reinforcement learning that can be found in the computer-science and economics literatures (e.g., Erev and Roth 1998, Sutton and Barto 1998). Nor have we considered the use of various combinations of the current models.

Nevertheless, there does exist one model that we *have* analyzed but do not include in the body of the paper. This is the so-called "probability-matching" model that has a long history within the research literature on human and animal choice (e.g., Robbins and Warner 1973).

The reason we do not explicitly include the results is twofold. First, the model is not directly applicable to the bandit setting. Second, when a natural variant of the model (which is applicable to the setting) is fit to subject data, it performs quite poorly. For more on the model and results, see Appendix J (online).[6]

To better understand how well these models apply to the bandit problem, we conduct two sets of tests. The first considers a general prediction consistent across all models: That the expected number of consecutive trials on an arm is increasing and convex in the average quality (reward) of the arm. The second is a more detailed discrimination among the various models, based on the choices that they recommend for a given observed history. Before we provide the details of the tests, we describe the experimental setup.

4. Experimental Setup and Preliminary Results

To empirically test the models of customer response, we have developed and run experiments in which subjects face the bandit problems described above. The experiments are designed to allow us to perform the more general test for convexity, as well as the more detailed, trial-by-trial analysis of each model's consistency with subjects' observed choices.

In this section we describe, in detail, how the experiments are structured. We also provide some descriptive statistics that give a general sense of the subjects' performance in the experiments.

4.1. The Value Distribution

The models we test are, ultimately, meant to represent customer switching due to variation in product or service quality. The hedonic nature of these experiences is difficult to control or monitor, however. For example, the same physical measures of service quality—such as speed, accuracy, or courtesy—may be perceived or valued differently by different people.

Therefore, we operationalize differences in perceived value with money; that is, the dollar reward received when sampling an arm is a proxy for the value received. This payment procedure is standard practice in experimental economics and is intended to induce participants to take their task seriously because real money is at stake (Friedman and Sunder 1994).

We define the payout distributions of the arms to be Bernoulli random variables: With probabil-

[6] To interpret the appendix's Figure 22, first read §6.

ity $\mathsf{P}\{i \text{ wins}\}$, arm i pays \$0.10, and with probability $\mathsf{P}\{i \text{ loses}\} = 1 - \mathsf{P}\{i \text{ wins}\}$, it pays nothing. Note that the shifting or scaling of outcomes $\{\$0.00, \$0.10\}$ does not affect the choices the models recommend.[7] Furthermore, the use of two-point payout distributions allows us to avoid problems associated with subject utilities that may vary nonlinearly with outcomes, controlling for risk preferences that vary across subjects.

One concern about the magnitude of the payments is the steepness or flatness of the resulting curve of total rewards obtained by subjects (Harrison 1989). In addressing this issue, we considered a spectrum of alternatives, from running hypothetical experiments (as in previous research), in which the reward function is everywhere zero, to those in which individuals make very few, high-stakes choices. Previous research has demonstrated that the move from hypothetical decisions (with no payment) to real decisions (with small payments) produces a significant change in subject behavior, while the move from small payments to large payments does not significantly affect behavior (Camerer and Hogarth 1999).

Thus, our payment scheme has sought to balance the need for payoffs that are responsive to subject choices with that for collecting enough data to adequately distinguish among models. The \$0.10 per win reflects a small but significant reward for making careful choices, and it has allowed us to collect ample data to fit our models (347 choices per subject). Furthermore, a number of descriptive statistics (reported below) suggest that most participants understood the experiment and paid attention to the task at hand.

4.2. The Number and Nature of Arms

Every subject plays a series of three two-armed Bernoulli bandit problems. We will sometimes refer to each problem as one of three "sessions" in which a subject participates. Participants are informed that, in a given session, each of the two arms has a fixed, but unknown, probability of winning; that is, the probability of success may vary from arm to arm but will remain constant over time for an individual arm.

4.3. Prior Information

Before they begin, participants do not know the two arms' probabilities of winning. We do, however, show participants some prior information about the arms. Specifically, we report (in writing) that each of the two arms has been sampled from three times and that two of the three trials were successes.

Given this information, subjects know that the prior performance of the two arms is equivalent and that the arms have $\mathsf{P}\{\text{win}\} \in (0, 1)$, so they are not degenerate. By communicating that there were two successes in three prior trials, we explicitly provide subjects with the same prior data that we use when fitting the various models to observed behavior. For details on how we fit the models, see §6.

4.4. Discounting

Recall that the form of the Gittins index result (2) for rational subjects depends on an infinite horizon setting with constant discounting. In experiments, a common method operationalizing a constant discount rate is to generate a constant probability $(1-\alpha)$ that any given round of the experiment will be the last. In our experiment $\alpha = 0.99$, and our instructions communicated to subjects that there would be a 1% chance that any given trial would be the last (equivalently, a 99% chance that it would not be the last).

To determine the number of trials to be run, we then implemented the randomized procedure only once, *before* any of the experiments were run, and we used the same three sample outcomes for all subjects: 95, 117, and 135 rounds. In this fashion, the procedure ensured that results would be directly comparable across participants.

4.5. Subject Recruitment

Participants in the experiment were undergraduate and graduate students at a large university on the east coast of the United States. Signs were posted on campus, offering money for participating in an experiment. There were tear-off slips on the bottom of each sign containing the URL for the experiment, and highlighting that only university students were eligible.

A student arriving at the website was asked for a university ID number. Having entered the ID number, the student then proceeded through the experimental instructions and tasks. At the end of the experiment, the student's Web browser generated a receipt that

[7] Theoretically, we can normalize the reward from winning to one and that from losing to zero. The expected reward from choosing an arm is the probability of winning on that arm.

included his or her earnings from the experiment and the university ID number that was entered at the start of the session.

Each participant then printed this receipt and brought it to an on-campus office. He handed the receipt and ID, which displayed the ID number and a picture of the student, to an administrator who checked that the person presenting the receipt was the person pictured in the ID and that the two ID numbers matched. If everything matched, the student was paid the earnings listed on the receipt.

4.6. Experimental Treatments

The experimental design involved three treatments, one for each of the three bandit problems: One that required 117 choices between arms with $\mathsf{P}\{i \text{ wins}\}$ of 0.15 and 0.40; a second that required 95 choices between arms with $\mathsf{P}\{i \text{ wins}\}$ of 0.40 and 0.40; and a third that required 135 choices between arms with $\mathsf{P}\{i \text{ wins}\}$ of 0.40 and 0.65. This design allowed us to test hypotheses about convexity by comparing the aggregate frequency of choices of an arm, given its win rate.

Treatments were run within subject, so each participant saw all three treatments. This allowed us to test predictions at the individual level (because each participant saw all three pairs of arms). We used a random number generator within the Web page to randomly assign each participant to one of the six possible orders in which the three sessions could occur. Thus, roughly one-sixth of the subjects saw each of the possible treatment orderings.

4.7. Software-Based Implementation

The experiment was implemented via computer on a Web page that could be accessed by a typical browser. Participants were told the website's address, logged into the system, and played the game.

The bandit problems were implemented as the repeated choice between two colored decks of cards. In each of the three sessions; the success probabilities (0.15/0.40, 0.40/0.40, 0.40/0.65) and colors (red/blue, green/gray, yellow/purple) assigned to the two decks and their location on the screen (left and right) were randomly selected for each participant.

The decks were composed of cards that state "YOU WIN!" or "YOU LOSE!" and their compositions remained constant over time. Each time a subject chose one of the available decks, an animation played that showed the deck being shuffled, one card being chosen at random, and the outcome of the trial, win or lose. The card was then replaced, and the deck reverted to its initial state.

At all times, subjects also saw the balance of their winnings as it accumulated at $0.10 per win. Participants could click a "history" button, which would display summary statistics concerning total wins and losses, as well as the entire history of their choices and the resulting outcomes.

At the end of each of the three bandit problems, each subject answered a short questionnaire. Demographic and other information was collected once the entire game had ended. Appendix B (online) presents the entire set of instructions and participant views of the experiment.

All participants earned a $5 participation fee plus their accumulated earnings from the experiment. At the end of the experiment, subjects were informed of their total earnings and asked to print a receipt from their browsers. They brought the receipt to an assistant, who checked it against their ID, obtained their signature, and paid them their earnings.

4.8. Data Collected

There were 373 participants who logged onto the system. Of these, 227 completed the experiment. Thus, we have collected data on 347 $(95+117+135)$ choices times 227 participants $=78{,}769$ choices between two options. For each of the 227 participants, we have a complete record of each of three treatments of the bandit problem, as well as answers to end-of-treatment and end-of-experiment questions. An example of one session of a subject is shown in Figure 1.

4.9. Descriptive Statistics

Of the 227 subjects who completed the experiment, 32 had at least one session in which the last "run"—a set of consecutive trials on a given arm—is the only run on that arm. In these cases, the sole run on that arm is artificially truncated by the end of the experiment, and tests for the convexity of the average run length become difficult to perform. We have, therefore, excluded these subjects from our tests of convexity and of model fit, and our final data set includes 195 subjects, 347 trials per subject, for a grand total of 67,665 recorded choices.

Figure 1 **Results for Subject 32, Session 2**

	Left	Right		Left	Right		Left	Right
1	LOSE		46	WIN		91	LOSE	
2		WIN	47	LOSE		92		LOSE
3		WIN	48		WIN	93	WIN	
4		WIN	49		WIN	94		WIN
5	WIN		50		WIN	95		WIN
6		WIN	51		WIN	96		WIN
7	WIN		52		LOSE	97		WIN
8	LOSE		53	WIN		98		WIN
9		WIN	54		WIN	99		WIN
10		WIN	55		LOSE	100		LOSE
11		WIN	56	LOSE		101	WIN	
12		LOSE	57		LOSE	102		LOSE
13	LOSE		58		WIN	103	LOSE	
14		WIN	59		LOSE	104		WIN
15		LOSE	60	LOSE		105		WIN
16	WIN		61		LOSE	106		WIN
17		LOSE	62		WIN	107		WIN
18		WIN	63		WIN	108		WIN
19		LOSE	64		LOSE	109		WIN
20	LOSE		65		WIN	110		WIN
21		LOSE	66		WIN	111		LOSE
22		WIN	67		WIN	112		LOSE
23		WIN	68		WIN	113	LOSE	
24		WIN	69		WIN	114		WIN
25		WIN	70		WIN	115		WIN
26		WIN	71		LOSE	116		LOSE
27		WIN	72		LOSE	117		WIN
28		WIN	73	WIN		118	WIN	
29		LOSE	74		LOSE	119		WIN
30		LOSE	75	WIN		120		LOSE
31	WIN		76	LOSE		121	WIN	
32	LOSE		77		WIN	122		WIN
33		WIN	78		LOSE	123		WIN
34		WIN	79	WIN		124		LOSE
35		WIN	80		WIN	125		WIN
36		WIN	81	LOSE		126		WIN
37		WIN	82		LOSE	127		WIN
38		WIN	83		WIN	128		LOSE
39		WIN	84		WIN	129	WIN	
40		WIN	85		LOSE	130		LOSE
41		LOSE	86	LOSE		131	WIN	
42	LOSE		87		WIN	132		WIN
43		WIN	88		WIN	133	LOSE	
44		LOSE	89		LOSE	134		WIN
45		LOSE	90		LOSE	135		WIN

Note. P{left wins} = 0.40 and P{right wins} = 0.65.

The 195 subjects took, on average, 12 minutes and 53 seconds to complete all three sessions, or about 2.2 seconds per decision. As they moved from the first to the last session, subjects also took less time to make each decision. This may be due either to learning of the experimental setup or to boredom or fatigue. Subjects' payoffs from the sessions ranged between $17.50 and $23.50, with an average of $20.50.

Of the subjects, 65% ($n = 126$) were male, and 95% ($n = 185$) were undergraduates. Almost 90% ($n = 173$) were of caucasian or Asian origin. All were between 17 and 25 years of age. Analysis of variance and t-tests did not reveal significant differences in average winnings across these groups. For details on these demographic data, see Appendix C (online).

A number of statistics suggest that the variation in total rewards was sufficient to motivate participants to understand the experiment and pay attention to the task at hand. Only 20 of 195 subjects reported being confused by some aspect of the experiment. During the experiment, 56% of the subjects viewed the history screen (which summarized their previous choices and the outcomes) at some time, and those who viewed the history visited this screen an average of 9.4 times during the exercise. After each session, we also asked subjects which arm had the higher probability of reward: left, right, or neither. Overall, subjects answered correctly more than 69% of the time; and when the arms had different P{win}s, more than 80% answered correctly.[8]

We also considered how summary statistics varied across experimental conditions. For each subject, we calculated the average run length in a given session, the total number of trials (95, 117, or 135) divided by the total number of runs in that session, and compared average run lengths across the different arm conditions. When both arm probabilities were equal (0.40/0.40) the average run length was the smallest, at 7.6. Average run length was 8.6 when the choice was 0.15/0.40 and 13.3 when the choice was 0.40/0.65. On average, subjects also took slightly more time—about 0.1 seconds more per round—to complete the sessions when the two arms' probabilities were the same (0.40/0.40), rather than different (0.15/0.40 or 0.40/0.65). Thus, when the arms' winning probabilities were the same, so there was no clear best arm, there was more switching and subjects spent more time per trial.

[8] When P{left wins} = P{right wins} = 0.40, the fraction of correct answers—that the two arms had the same probability of reward—dropped to 37%. Of course, to say that the two arms had the same P{win} is a point prediction that is a less likely outcome than "left greater than right" or "right greater than left."

Longer run lengths and times per arm were also positively correlated with total earnings. Ordinary least squares (OLS) linear regression of subjects' earnings on average run length (defined as $\bar{\tau}$ in §5) and on average time per trial showed that a one-unit increase in average run length was associated with a \$0.07 average increase in total winnings, and a one-minute increase in total completion time (equivalently, a 0.173-second increase in time per round) was associated with a \$0.03 average increase in total winnings.[9] Of course, the fact that larger run lengths were associated with higher earnings may reflect that subjects tended to stay on "better" arms for longer runs.[10]

5. The Expected Switching Time

Suppose that at some arbitrary period, t, a subject has last sampled from arm i, and let τ be the number of additional periods that he or she will continue to sample from i, before switching to a competitor. Then asymptotic results suggest that the expected switching time, $\mathsf{E}[\tau]$ is increasing and convex in the average quality (reward) provided by i (see online Appendix D).

A test of this property is interesting to us for two reasons. First, it is a basic check of whether or not subjects' actual behavior corresponds to a prediction common to many models. Second, if true, the increasing-convex property is also consistent with a broader claim made in the popular management literature: that marginal increases in service quality can have increasingly dramatic benefits in customer loyalty and lifetime value (e.g., Jones and Sasser 1995). Therefore, our first set of tests focuses on $\mathsf{E}[\tau]$.

5.1. Estimation

In constructing tests for monotonicity and convexity, we control the quality of the arm and measure the resulting switching times. In the context of Bernoulli arms, $\mathsf{P}\{\text{win}\}$ is the measure of quality, and in the experiments, every subject plays three pairs of arms with the same probabilities of winning: 0.15 versus 0.4; 0.4 versus 0.4; and 0.65 versus 0.4. The quality points we have chosen are evenly spaced, with $0.65 - 0.40 = 0.40 - 0.15 = 0.25$, to facilitate testing for convexity. Furthermore, in each treatment, at least one arm has $\mathsf{P}\{\text{win}\} = 0.40$, so that the quality of the arm that is *not* used in the convexity test remains constant across treatments. In the 0.40 versus 0.40 treatment, either arm may be used in the convexity tests, and for this treatment, we report the results of both arms.

The estimate of $\mathsf{E}[\tau]$ is more difficult to calculate. One easily calculated measure is the average number of consecutive trials—or average "run length"—on each arm, calculated as

$$\bar{\tau}_i = \frac{\text{total number of trials on an arm}}{\text{total number of runs on an arm}}, \qquad (8)$$

where $i = l$ for the left arm and $i = r$ for the right.

We test that $\bar{\tau}$ is increasing and convex in $\mathsf{P}\{\text{win}\}$ as follows. Let $\bar{\tau}_{p,j}$ be the average run length for an arm with $\mathsf{P}\{\text{win}\} = p$ and subject j. For each subject, we first calculate the first difference between adjacent quality pairs: $\Delta_{1,j}(\bar{\tau}) = \bar{\tau}_{0.40,j} - \bar{\tau}_{0.15,j}$ and $\Delta_{2,j}(\bar{\tau}) = \bar{\tau}_{0.65,j} - \bar{\tau}_{0.40,j}$. Given the experimental setup, we know that these sample differences are independent across subjects, and we use the Wilcoxon signed rank test to check that the median of the differences is greater than zero (Lehmann 1975). Our test for convexity runs along the same lines. Here, we define each subject's second difference as $\Delta^2_j(\bar{\tau}) = \Delta_{2,j}(\bar{\tau}) - \Delta_{1,j}(\bar{\tau})$, and we check that it is positive.

Note that there exist two potential problems associated with using $\bar{\tau}_i$ as an estimate of $\mathsf{E}[\tau]$. First, subjects' last runs are censored. For example, from Figure 1, we see that Subject 32's last run consists of trials 134 and 135 on the right arm. At this point, the experimental treatment was ended. Had the subject been allowed to continue sampling, the run length might have been longer than two, however. A second potential problem concerns the possibility that the sequence of run lengths is not stationary. For example, the sequence of run lengths may be (stochastically) increasing or decreasing, rather than stationary.

[9] $R^2 = 0.15$ in the regression, and both coefficients were significant at the 0.05 level.

[10] We have also calculated descriptive statistics for the subjects who were excluded from our main analyses. The 32 subjects who completed the experiment (but had a single, truncated run on a given arm) spent about four minutes less total time (about 0.7 seconds less per decision) on average and won approximately the same amount as other subjects. Subjects who did not finish the experiment spent about 0.2 seconds more time per decision and won about \$0.0015 less per decision than the 195 subjects included in our analyses.

Indeed, our results suggest that over the course of a session there is a mild increase in run lengths.[11]

The overall impact of these effects is not immediately clear to us. On the one hand, the length of the censored run is longer than what was recorded. In this sense, censoring biases the average downward. On the other hand, there is an inspection bias present: Runs that are censored are likely to be longer than average. Even if run lengths are not stationary, to the extent that τ is *stochastically* increasing the quality of the arm, then these sample averages should be increasing in quality as well. Nevertheless, there are modifications that can correct for the problem.

One simple alternative, which we denote $\hat{\tau}$, eliminates the last, truncated run from the calculation of the sample average run length. As with the $\bar{\tau}_{p,j}$'s, we can use subjects' $\hat{\tau}_{p,j}$'s to calculate $\Delta_{1,j}(\hat{\tau})$, $\Delta_{2,j}(\hat{\tau})$, and $\Delta_j^2(\hat{\tau})$ and use the Wilcoxon signed rank test to check for the increasing and convex properties.

A second alternative tests for convexity of a *single* run, starting in a given trial of the session, and we measure the length of the run that begins at Trial 1, which we call τ_1. By choosing the first round, we ensure that runs are not truncated. (The only subjects whose first runs are truncated are those that never change arms.) The choice of the first round also ensures that all subjects have exactly the same information about the arms as the run begins, so that differences among run lengths do not result from experience or informational differences.

In many respects, the second alternative is preferable. It has the drawback, however, of not allowing for the within-subject comparisons across the three arms' qualities. This is because, given our randomization scheme, the success probability of the arm first chosen by a subject is not controllable. When faced with 0.15/0.4 or 0.4/0.65 treatments, many subjects (unknowingly) first chose the arm that had a probability of winning of 0.40.

5.2. Results

While we have calculated relevant statistics for all three measures of run length, Figure 2 displays confidence intervals only for $\bar{\tau}$ and τ_1. Because the results for $\hat{\tau}$ are similar to those for $\bar{\tau}$, we omit their graphical display.

The figure's left panel shows confidence intervals for $\bar{\tau}_p = (1/n)\sum_{j=1}^{n} \bar{\tau}_{p,j}$, $p \in \{0.15, 0.40, 0.65\}$. Each interval shown in the panel is calculated as a sample average ±2 times the standard error of the estimate of the mean. In each of the three intervals, the reported arm competes against an arm with $\mathsf{P}\{\text{win}\} = 0.40$, and in the case of 0.40 versus 0.40 treatment, we report the results of both arms.

Here, the sample averages are clearly increasing with increasing differences, and the confidence intervals do not overlap—a further indication that the increasing property holds. Wilcoxon signed rank tests for first and second differences are all vanishingly small (reported as 0 in S-plus).

For $\hat{\tau}$, the confidence intervals are qualitatively the same as for $\bar{\tau}$ (see Appendix E, which is online). Similarly, the p-values of the Wilcoxon signed rank tests were also negligible. Thus, the results are consistent with hypothesis that $\bar{\tau}$ and $\hat{\tau}$ are both increasing and convex in $\mathsf{P}\{\text{win}\}$.

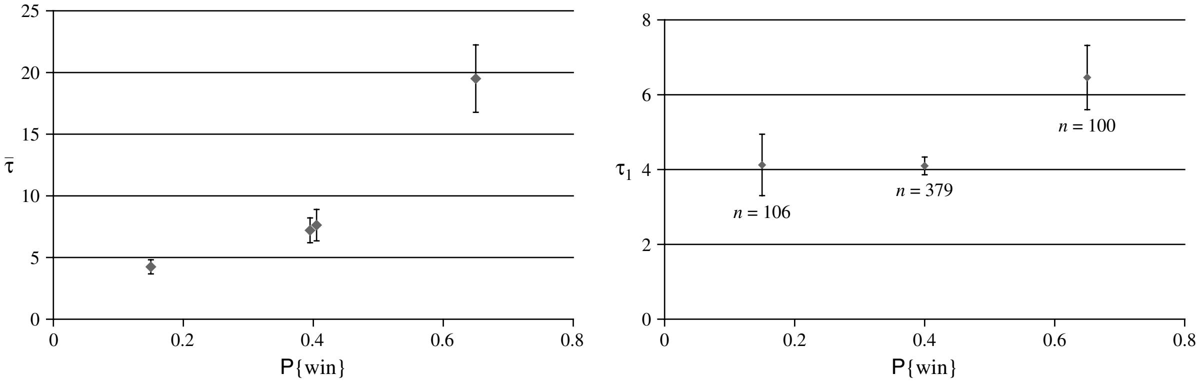
The right panel of Figure 2 shows the results for the single run that starts at the first trial. Here, we do not have paired data across the three treatment conditions. Furthermore, the numbers of subjects for which we have results varies across the conditions: For $\mathsf{P}\{\text{win}\} = 0.15$ the number of subjects is $n = 106$; for 0.40, $n = 379$; and for 0.65, $n = 100$. In this case, the second difference in the point estimates of the average length appears to increasing, though the first difference is *not* increasing from a $\mathsf{P}\{\text{win}\}$ of 0.15 to 0.40. Furthermore, the relatively smaller sample sizes for $\mathsf{P}\{\text{win}\}$ of 0.15 and 0.65 result in confidence intervals that are much wider than that for $\mathsf{P}\{\text{win}\}$ of 0.40.

Although the lack of pairing of the data makes it more difficult to test for an increases in the second difference, we can use a Mann-Whitney test (also known as the Wilcoxon rank sum test; see Lehmann 1975) to check whether or not the medians are increasing from one treatment condition to the next. While τ_1 is not significantly increasing from 0.15 to 0.40, it is increasing from 0.40 to 0.65, as well as over the whole range, 0.15 to 0.65.[12]

[11] On average, the number of switches decreased from 12.65 in the first half of each session to 9.97 in the second half, a reduction of about 21%.

[12] The p-values were as follows: 0.34 for τ_1 increasing as $\mathsf{P}\{\text{win}\}$ increases from 0.15 to 0.40; and 0.02 for τ_1 increasing as $\mathsf{P}\{\text{win}\}$ increases from 0.40 to 0.65 and from 0.15 to 0.65.

Figure 2 Convexity Results for the Average Run Length



In summary, the results of this section are generally consistent with the hypothesis that $\mathsf{E}[\tau]$ is increasing and convex in the quality of an arm. Of course, because the observed behavior is consistent with the prediction does not prove that run length is increasing and convex. We will return to this point in §7.

6. Estimating the Models' Fit to the Data

In this section, we consider the more difficult question of distinguishing among the various models' fits to the data. We begin the section by describing how the models of §3 are applied to the experimental data and how we calculate the indices associated with each of the models. We then describe our estimation procedures in fitting subjects' choices. Finally, we compare how well the models fit subjects' actual choices.

Our results show that the ES and HH models provide the best in-sample fits, both on an aggregate and subject-by-subject basis. We also show, however, that the HH model's fit may be an artifact of subjects' long-run lengths. Among the hierarchy of Gittins index-derived models, the Simple model performed best. Thus, there exist models that provide *both* analytical tractability and a reasonable fit to subjects' observed choices.

6.1. Calculating the Models' Indices

We first describe how the models' indices are calculated in the context of the experiment's Bernoulli outcomes.

6.1.1. Gittins Index Model. For the Gittins index model (2), we use results from Gittins (1989) that explicitly calculate Gittins indices, G, for Bernoulli bandits with conjugate (beta-distributed) priors. Specifically, given a beta-distributed prior, a discount rate, α, and numbers of wins and losses prior to t, ω_{t-1}^{i} and ℓ_{t-1}^{i}, the index for an arm is a function of the triple $(\alpha, \omega_{t-1}^{i}, \ell_{t-1}^{i})$. Table 11 in Gittins (1989) lists $G(\alpha, \omega_{t-1}^{i}, \ell_{t-1}^{i})$ for arms in which $\alpha = 0.99$ and ω_{t-1}^{i} and ℓ_{t-1}^{i} range from 1 to 40.[13]

In calculating each arm's Gittins index for any experimental trial, we assume the prior is beta distributed and use the results from Gittins (1989). Given the prior information we communicate to subjects—that $\alpha = 0.99$ and that both arms had won twice and lost once in three prior trials—we further assume that $(\alpha = 0.99,\ \omega_0^{i} = 2,\ \ell_0^{i} = 1)$. After each choice of an arm and outcome, that arm's ω_t or ℓ_t is updated and its G recalculated.

Note that the use of beta-distributed priors is purely for computational tractability, and in the experimental sessions, we did not inform subjects that the prior distribution was of this form. Thus, our experiment does *not* provide a sharp test of whether or not subjects are rational. Rather, it tests how well the "rational model" with the given beta prior fits subjects' choices. This is in keeping with our original aim of validating models for use in the context of competitive analysis. To emphasize the distinction between

[13] For ω_{t-1}^{i} or ℓ_{t-1}^{i} greater than 40, Gittins (1989) provides an approximation—Equation (7.16), fitted with parameters from Tables 12–14—that allows for calculation of indices that are typically precise within four decimal places.

a test for rationality and a test for model fit, we call the model "Gittins" (rather than "rational") when we report our experimental results.

Furthermore, even if we had informed subjects, in words or through pictures, that the prior is of a beta ($\omega_0^i = 2$, $\ell_0^i = 1$) form, neither the prior distribution nor the discount rate *implicitly* used by the subject is observable. In theory, we might estimate ω_0^i, ℓ_0^i, and α to accommodate subjects' unobserved, idiosyncratic priors and discount rates. The result would be a generalized model with three free parameters, and we would search over all feasible triplets $(\alpha, \omega_0^i, \ell_0^i)$ to find the initial parameters that generate the least inconsistency between model and experimental data. In practice, however, the calculation of the Gittins index is burdensome in and of itself, and we have not attempted to search among this broader class of Gittins index policies.[14]

6.1.2. Myopic Model. For the Myopic model we, again, use beta priors for convenience, and the resulting calculations are straightforward. Specifically, (3) does not require a discount rate, α, and given a beta prior, the index is a straightforward function of previous wins and losses:

$$M(\omega_{t-1}^i, \ell_{t-1}^i) = \frac{\omega_{t-1}^i}{\omega_{t-1}^i + \ell_{t-1}^i}. \tag{9}$$

Given this form, it is not difficult to incorporate the effect of changes in the prior distribution on the indices.

Therefore, we test two versions of the Myopic model. The first includes no free parameters and assumes ($\omega_0^i = 2$, $\ell_0^i = 1$). We call this the "Myopic-0" model. The second incorporates one free parameter. Specifically, we define a common ω_0^i for both arms and let it range between 0.01 and 2.99, in increments of 0.01. We then define a common $\ell_0^i = 3.0 - \omega_0^i$. Thus, the second version still requires that both arms have the same initial prior, but it allows the shape of the prior to vary. We call this model "Myopic-1."

Two elements of the parameter range are worth noting. First, to be consistent with the prior win-loss information we report to subjects (two wins in three prior trials), we bound ω_0^i away from 0 and 3, which respectively reflect beliefs that the probability of a win is zero and one. Second, by requiring $\omega_0^i + \ell_0^i = 3.0$, we fix the "strength" of the initial prior to be consistent with the quantity of prior information we report to subjects: the results of three prior trials.[15]

6.1.3. Simple Model. Recall that the Simple model hypothesizes that customers think of arms as being good or bad, with expected rewards of μ_G and μ_B. Algebraic manipulation (provided in Appendix F, which is online) demonstrates that we can write the index for arm i at time t as

$$\tilde{S}(\omega_{t-1}^i, \ell_{t-1}^i) = \omega_{t-1}^i \cdot \mathcal{U} - \ell_{t-1}^i \cdot \mathcal{D}, \tag{10}$$

the result of a series of ω_{t-1}^i "up steps" ($\mathcal{U}$) and ℓ_{t-1}^i "down steps" ($\mathcal{D}$). Furthermore, without loss of generality, we can normalize $\mathcal{U} \equiv 1$. The result is a model with one free parameter, $\mathcal{D}$.

For arm i with probability μ_i (not necessarily μ_G or μ_B) of winning, the expected "drift" of the random walk $\tilde{S}$ at time t equals $\mu_i \mathcal{U} - (1 - \mu_i)\mathcal{D}$. Then given $\mathcal{U} \equiv 1$ and some fixed $\mathcal{D}$, the subject's "aspiration level"—the average quality level required of i so that the drift is nonnegative—is

$$\mu^* = \frac{\mathcal{D}}{\mathcal{D} + 1},$$

and once a Simple subject begins sampling from an arm with $\mu_i \geq \mu^*$, his or her expected switching time is infinite (see the discussion in §3.2, as well as Gans 2002b). When estimating the fit of subjects' choices with the Simple model, we systematically vary $\mathcal{D}$ so that μ^* ranges from to 0.0033 to 0.99 in even increments of 0.0033.

We bound $\mathcal{D}$ away from zero so that the model is required to penalize arms for bad outcomes. In contrast, for $\mathcal{D} = 0$, the Simple model recommends an arm with the greater number of wins—without regard to numbers of losses—and would provide (perhaps unfairly) good fits for subjects that never change arms, no matter how many losses. (For example, see the results for HH-n in §6.3.2, below.)

[14] There exist closed-form approximations to the Gittins index that could be used to fit idiosyncratic subject prior and discount-rate information (Chang and Lai 1987, Brezzi and Lai 2002). The expressions are not accurate for discount rates that are significantly less than one, however.

[15] This setup implicitly assumes that, before being informed of two successes in three prior trials, subjects have noninformative priors. An alternative, which we have not tested, would be to let $\omega_0^i = \gamma + 2$ and $\ell_0^i = \gamma + 1$ for both arms, where $\gamma > -1$ is the model's single free parameter.

6.1.4. Last n. As with the previous models, for convenience we use beta prior, so at trial t the index for arm i is the ratio of the number of wins and losses in the previous n trials on the arm. Formally, we let $\omega_t^i(n)$ and $\ell_t^i(n)$ be the number of wins and losses in the last n trials on arm i, so that $\omega_t^i(n)+\ell_t^i(n)=n$, and

$$L(\omega_{t-1}^i(n), \ell_{t-1}^i(n)) = \frac{\omega_{t-1}^i(n)}{\omega_{t-1}^i(n)+\ell_{t-1}^i(n)}. \qquad (11)$$

While we have tested models for $n=1,2,3,4,5$, we report results only for $n=1,3,5$. This allows the figures to be less cluttered and easier to read, and the omitted results (for $n=2,4$) are consistent with the broader trends seen across $n=1,3,5$. We have also tested a "meta" model that treats n as an additional free parameter and uses the n associated with each subject's lowest Bayesian Information Criterion (BIC) score. The results of this meta-model are not fundamentally better than those of the basic Last-n family.

Finally, we note that, at $t=0$ the Last-n model requires data for periods $t=-1,-2,\ldots,-n$. For simplicity, we have initialized the record of all these prior outcomes to be wins. While this assumption is not consistent with the prior information shown to the subjects, it is consistent with the initial conditions required for fitting the HH family of models (see below). As we will see in the next subsection, differences in fit are substantial across models, and we do not believe that these initial conditions have significantly affected our results.

6.1.5. Hot Hand. Recall that the HH family of models uses only the results of the most recently sampled arm to decide which arm to sample next. Specifically, if the n previous trials on the current arm were all losses, then the HH-n model recommends switching to the other arm; otherwise, the model recommends continuing to sample from the current arm.

While the HH rule is *not* index based, when fitting the model to subjects' observed choices, it will be convenient for us to define it as an index rule. Therefore, we formally define HH-n indices as follows. For a subject that sampled from arm i at time $(t-1)$, we define the indices for arms i and $j\neq i$ to be

$$\begin{aligned} HH^i(\ell_{t-1}^i(n)) &= 1\{\ell_{t-1}^i(n)<n\} \quad \text{and} \\ HH^j(\ell_{t-1}^i(n)) &= 1-HH^i(\ell_{t-1}^i(n)), \end{aligned} \qquad (12)$$

where $\ell_{t-1}^i(n)$ denotes the number of losses in the previous n contiguous trials on arm i. The rule then recommends choosing the arm with the larger of the two indices.

If at time t a subject chooses an arm, i, that she had chosen in the previous trial, then $\ell_t^i(n) = 1\{i \text{ loses at } t\} + \ell_{t-1}^i(n)$. Because the counter $\ell_{t-1}^i(n)$ only tracks the number of losses in the previous n *contiguous* trials on arm i, we reset $\ell_{t-1}^i(n)=\ell_{t-1}^j(n)=0$ whenever the subject switches arms: that is, whenever $HH^i(\ell_{t-1}^i(n)) \neq HH^i(\ell_t^i(n))$. This implies that HH-$n$ recommends staying on the current arm whenever the current run on an arm is less than n trials.

Because the prior information we provide to subjects at the start of each session does not distinguish the order in which "prior" samples of the two arms were made, it is not well determined whether a subject's first trial in a given session represents a switch to a new arm or the continuation of a run on the current arm. Therefore, for simplicity, we assume that the first trial represents a switch, and we reset the associated loss counters, $\ell_0^i(n)=\ell_0^j(n)=0$, accordingly.

6.1.6. Exponential Smoothing. The ES model (7) is implemented in a straightforward fashion. We test two versions of it. The first model (ES-1) has the smoothing weight, γ, as its one free parameter. The initial index of each arm is fixed at $ES_0^i=2/3$, so that it matches the win-loss ratio reported as prior information. Note that, for $\gamma=1$, ES-1 corresponds to a Last-1 model, so to better distinguish between the two models, we bound γ away from 1. Similarly, for $\gamma=0$, both arm's indices would equal $2/3$ for all t and would not be informative. Therefore, we vary γ from 0.01 to 0.99 in increments of 0.01.

The second model (ES-2) also treats the initial index ES_0^i as a free parameter. In this case, we vary both γ and ES_0^i from 0.01 to 0.99 in increments of 0.01, for a total of roughly 10,000 possible combinations of free parameter values.

6.2. Fitting the Models to Subjects' Choices

Let the generic index, I, represent the index of the model being used, and suppose that, at trial t in a given session, a model's indices are I_t^l and I_t^r for the left and right arms. If $I_t^l > I_t^r$, then a subject whose choices are dictated by the model will choose the left arm. In fact, at t, the subject will either choose the left arm or not, and a straightforward and readily observable measure of consistency would simply record whether or not the subject's choice matches the prediction.

An aggregate measure of consistency over all 347 trials would then be the total number of incorrect model predictions, the smaller the number the better. The determination of this number is trivial for the Gittins, Myopic-0, Last-*n*, and HH-*n* models because they have no free parameters. For the Myopic-1, Simple, ES-1, and ES-2 models, one can search for the values of the respective free parameters that minimize the total number of errors across all 347 trials.

Alternatively, at any given trial, one might further judge *how well* a subject's choice matches that prescribed by a model, rather than simply whether or not the observed choice is consistent. A common means of judging the degree of consistency is through the use of random utility models (Anderson et al. 1992). Here, one posits that the value of a left or right choice at a particular trial is randomly distributed and that only the mean of the distribution is captured by the model indices I_t^l and I_t^r. In the context of our experiments, the random fluctuation might be ascribed to errors (or "trembles") in judging the value of the choice.

If, from trial to trial, this random noise is independently and identically distributed according to a Gumble (double exponential) distribution, then we have a so-called logit model. In this case, the probability of choosing the left arm at trial t is

$$\mathsf{P}\{\text{choose left}\} = \frac{e^{\beta I_t^l}}{e^{\beta I_t^l} + e^{\beta I_t^r}}. \qquad (13)$$

Similarly, $\mathsf{P}\{\text{choose right}\} = 1 - \mathsf{P}\{\text{choose left}\}$.[16]

Note that, by nesting the original choice models within the logit framework, we have imposed an additional (and unobservable) level of complexity, as well as the addition of a free parameter, β. The benefit is that we have a means of judging how well each observed choice matches a model's prediction. Furthermore, that measure is a probability, and we can easily add the 347 choices' log-probabilities to calculate an aggregate log-likelihood (LL) of observing each subject's outcome.[17]

A likelihood measure of consistency is particularly appealing in that it also lets us naturally correct for differences in the numbers of free parameters used by the models, something which is not easily accomplished when counting the total numbers of errors. We account for free parameters using the BIC; $\text{BIC} = -2\text{LL} + \text{d.f.} \times \ln(347)$, where d.f. is the numbers of degrees of freedom used by the model's free parameters. Indeed, BIC scores have the appealing property that they can be used to approximate Bayes factors, the posterior odds the data being explained by one nonnested model, rather than another (Kass and Raftery 1995).

In contrast, a straightforward count of the number of inconsistent choices can become problematic if the index takes on only a few values. For example, in the extreme case that a choice model has $I_t^l = I_t^r$, regardless of a subject's observed choices and outcomes, the number of inconsistent choices is always equal to zero, and the model "perfectly" fits the observed data. In contrast, the likelihood approach would record that each choice is completely random ($\mathsf{P}\{\text{choose left}\} = \mathsf{P}\{\text{choose left}\} = 0.5$, without regard to β), and the associated BIC score would, more appropriately, suffer. For this reason, we emphasize results obtained from using the logit model.[18]

The calculation of each model's LL requires the solution of a nonlinear optimization problem. For models with no free parameter, the LL is concave in β, and the optimal β can be found using standard optimization software. While the LL's of the Myopic-1, Simple, and ES models are also concave in β, they are not necessarily jointly concave in β and their free parameters. Therefore, for these models, we perform a nested optimization procedure: The top level is a

[16] Thus, the probability a subject chooses an arm increases in the index associated with that choice and decreases in the other index. As Allison (1982) has shown, this procedure is equivalent to that for estimating a discrete-time hazard model with a logistic regression function; that is, the likelihood scores generated by these models are equivalent. (See the section on discrete-time methods, particularly the discussion surrounding Equation (22).)

[17] Our implementation of the HH model defines the index of the recommended arm as one and that of the other arm as zero. While this choice appears to be somewhat arbitrary, we note that it is without loss of generality, since the β of the superimposed logit model acts as an independent scaling factor. Therefore, the LL derived from the HH model is independent of the scale of its indices. (Thanks to Ed Kaplan for pointing this out.)

[18] For more information on this effect, see Appendix G (online).

grid search over each model's original free parameters; then, for each of these grid values, we use a solver to find the optimal β and LL. We then record the best LL among all of the top-level grid values.

We implement this procedure for each subject to find one best set of free parameter values across all 347 trials. For each subject, we use the same free-parameter values across all three sessions. This approach reflects our assumption that each subject's parameters should be stable across treatment conditions, and it reduces the possibility of model overfitting.

Given the large quantities of data that we analyze, the use of BIC scores is also computationally appealing. In total, we have calculated more than 2 million scores for model fit—more than 10,500 free-parameter values associated with the various models for each of 195 subjects—and this task has been facilitated by the straightforward optimization required by LL and BIC scores.

6.3. Results

For each of 195 subjects, we have used the optimization procedure described above to find free parameters that minimize each model's BIC score for each subject, and we report the fit results below. Because of the numerous models considered, we report the results of the various models by family: first those that are derived from the Gittins index, then the Last-*n* group, then HH models, and finally ES.

6.3.1. Models Derived from the Gittins Index. Figure 3 presents two pictures of the BIC results for the Gittins index family of models. The left panel displays an aggregate, across-subject view of the models' performance, and the right panel shows the results of within-subject rankings of BIC scores.

More specifically, the curves in the figure's left panel represent cumulative distributions of BIC scores.[19] From the plot, we see that Simple and Myopic-1, the models with an extra free parameter, nearly dominate Myopic-0 and Gittins, even after the BIC score penalizes these models for the added parameter. Careful inspection shows that the Simple model does not, in fact, dominate the rest of the models in the 310-to-340 and the above-440 range. In the above-480 range, where the absolute fit of any of the four models is very poor, Myopic-0 outperforms Simple and Myopic-1 because its BIC score is not penalized for an extra free parameter.

The total BIC scores shown in the left panel's inset are consistent with these curves. When individual BIC scores are summed across individuals, we find that, in aggregate, the Simple model has the best performance, then Myopic-1, then Myopic-0, and finally the Gittins index itself.

We can also judge the quality of fit at the individual level. For every subject, we rank each model's fit by BIC score. For each model, the first shaded bar is the number of subjects for whom it was the first-best fit. The next white part is the number of subjects for whom it was the second-best fit, and so on. The right panel of Figure 3 displays these results.[20]

The results show that, on a subject-by-subject basis, the Simple model fits the observed choices best most often, in about 41% of all subjects. Again, the Simple model has roughly twice the number of first place fits when compared to Myopic-0 or Myopic-1. In this case, however, the Myopic-0 model slightly outperforms Myopic-1.

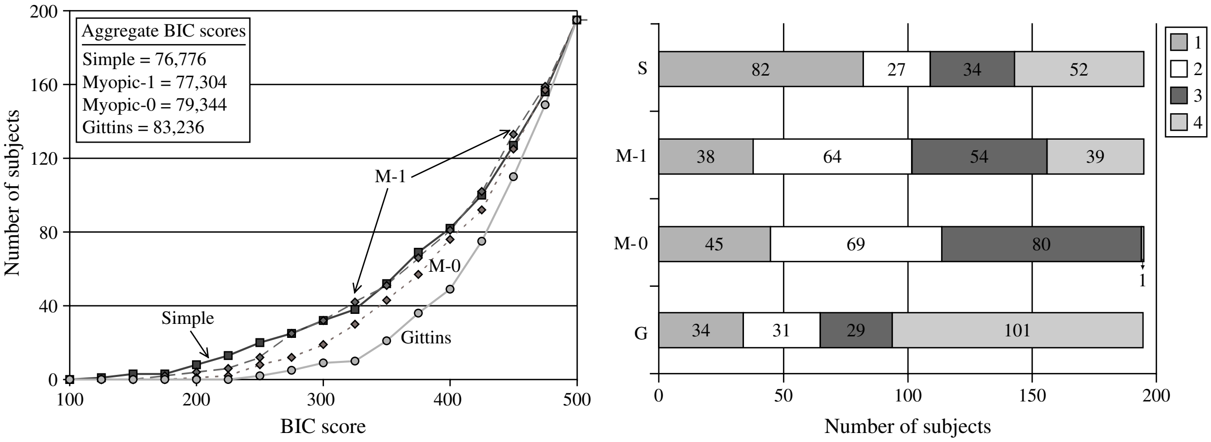
We also performed pairwise comparisons of the models across subjects using paired-*t* and Wilcoxon signed rank tests. In all but one case, the differences were significant, with desired *p*-values vanishingly small. For the one-sided (alternative) hypothesis that the Myopic-1 BIC > Simple BIC, the results were somewhat less strong, however. Here, the *t*-test resulted in *p*-value of 0.08, while the signed rank test resulted in a *p*-value of 0.11.[21]

[19] To derive a given model's curve we sort its 195 BIC scores from lowest to highest—that is, best to worst. Then we plot the number of subjects with BIC scores less that or equal to each value listed on the *x*-axis. Models whose curves are farther "up and to the left" are interpreted as better fitting the observed data.

[20] We note that, in the figure, a tie for equal BIC scores results in all relevant policies receiving the higher rank. Thus, while each policy's rankings add up to $n = 195$, adding up a given ranking (e.g., first place) across models may lead to a total that exceeds or falls short of 195.

[21] Of the 195 subjects included in the test, there was one extreme outlier, Subject 127. After removing this outlier from the data, the *p*-value for the *t*-test improved to 0.03, and that for the signed rank test to 0.08.

Figure 3 **BIC Scores for Models Derived from the Gittins Index**



6.3.2. Other Choice Models. We next present analogous results for the Last-*n*, HH, and ES models. To facilitate comparison with the Gittins index family of models described above, the figures include results for the Simple model as well.

Last n. The left panel of Figure 4 shows that the BIC scores associated with the Last-*n* family of models tend to be higher than those associated with the Simple model. Both the cumulative distributions and the aggregate BIC scores of the Last-5 and Last-3 models are dominated by the Simple model's scores. The Last-1 model outperforms the Simple model only when BIC scores are high and none of the models fits particularly well.

In contrast, the figure's right panel shows that, on a subject-by-subject basis, Last-1 appears to perform nearly as well as the Simple model. While a one-sided paired *t*-test, with an (alternative) hypothesis that Last-1 BIC > Simple BIC, was significant at nearly the 0.01 level, an analogous Wilcoxon signed rank test returned a weaker *p*-value, of 0.111.

As the left panel suggests, however, subjects with lower Last-1 BIC scores are those for whom neither model fits very well. For example, for the 88 subjects for whom the Simple model ranked first, the average Simple model's BIC score was 328.7, while for the 79 subjects for whom the Last-1 model ranked first, the analogous score was 392.4. In fact, in cumulative distribution of the first-ranking BIC scores, the Simple model also dominates Last-1.

Hot Hand. Figure 5 details the BIC scores of the HH family of models. One sees that, in contrast to the Last-*n* model, the HH models' fit to subjects' choices improves with larger *n*. HH-5, in particular,

Figure 4 **BIC Scores for Last-*n* Model**

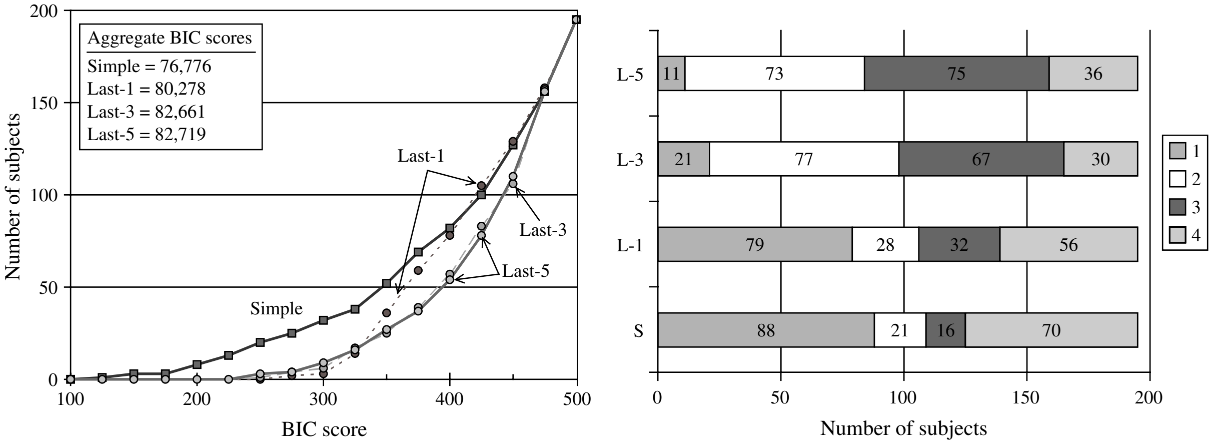
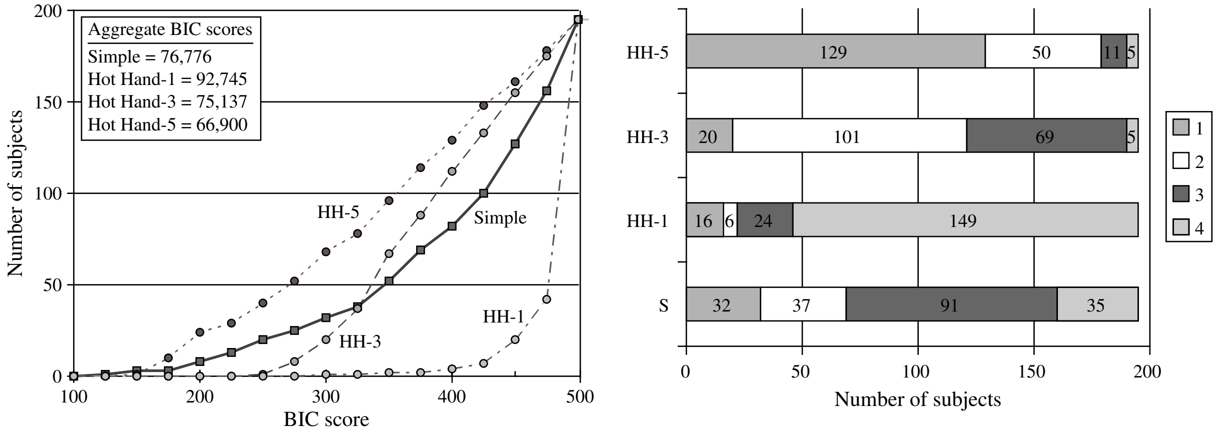


Figure 5 BIC Scores for the HH Model



appears to perform extremely well on both an aggregate and an individual basis. The left panel shows that the model's aggregate BIC score dominates that of the other models, as does the cumulative distribution of its BIC scores. The right panel shows that, on a subject-by-subject basis, HH-5 significantly outperforms the Simple model as well. Both a one-sided paired *t*-test and a Wilcoxon signed rank test that compare HH-5 and Simple BIC scores returned vanishingly small *p*-values.

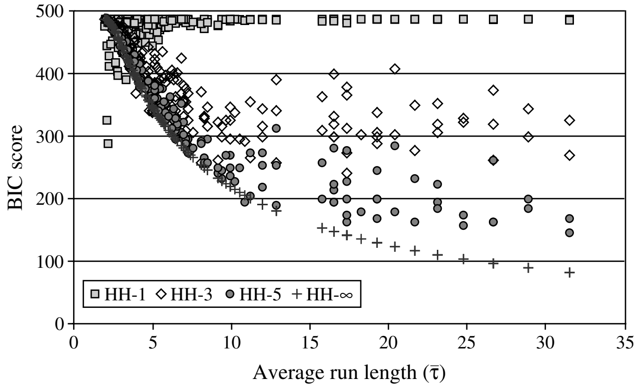
Note also that, within the HH family, the performance of the models is well ordered. In both of Figure 5's panels, HH-5 outperforms HH-3, and HH-3 outperforms HH-1. Indeed, while we do not report detailed results here, the results for HH-4 and HH-2 also fit within this ordering.

To better understand why the rankings appear to be ordered with n, we compared the models' BIC scores to subjects' average run lengths. The results, shown in Figure 6, are informative: the HH-5 model's BIC scores are strongly associated with average run length, with a nearly linear relationship between BIC and $\bar{\tau}$ for $\bar{\tau} \leq 10$.[22] Close inspection shows that a more attenuated form of the same type of relationship also holds for HH-3, and plots that include the results for the HH-2 and HH-4 models show a consistent progression: The larger the n, the more strongly average run length is associated with low BIC scores. This leads us to question the quality of the HH model's fit to the data.

More generally, consider a sample path in which an arm with a probability of p of winning is pulled over and over again. Then it can be shown that the expected fraction of trials in which a HH-n model recommends switching is of order $O((1-p)^n)$ (see Appendix D, which is online). As $n \to \infty$, the frequency of recommended switches approaches zero exponentially quickly.

Now suppose that in T trials, a subject switches arms m times. If an HH-n with very large n is fit to the data, it will register roughly m inconsistencies, perhaps fewer. In turn, if m is small then the fit associated with the HH-n model will be very good, no

Figure 6 Relationship Between BIC Scores and $\bar{\tau}$ in the HH Model



[22] To quantify the strength of the relationship, we also created a binary variable that took on a value of 1 for a subject if HH-5 was ranked first and zero otherwise. Tests for the probability of HH-5 obtaining a first ranking showed that $\bar{\tau}$ has, in fact, a significant positive impact. In a logit regression with an intercept, the coefficient for $\bar{\tau}$ was positive with a t-statistic greater than 4.6. A similar probit regression returned the same direction and significance.

matter what the relative winning probabilities of the two arms. Therefore, in a significant sense, the HH-n model may be obtaining good fits simply by picking up long run lengths, and we suspect that the BIC scores of HH-6, HH-7, etc. would keep improving.

As a Simple test of this hypothesis, we tested a HH-∞ model that recommends *never* switching. Its aggregate BIC score was 61,659, much better than that of HH-5. Furthermore, as the plot of BIC on $\bar{\tau}$ in Figure 6 confirms, HH-∞ displays a more extreme version of the pattern seen in the other HH models. Thus, we believe that the low BIC scores obtained by HH-5 are largely artifacts of subjects' long run lengths.

In contrast, there is some evidence against this phenomenon holding for the other models. Analogous plots for the other models do not show the same strong relationship between BIC scores and $\bar{\tau}$ (see Appendix H, which is online). Rather, BIC scores tend to be higher for small average run lengths and then drop for average run lengths above $\bar{\tau} \approx 7$. Above this cutoff, there appears to be only a mild, negative relationship between the two. An exception to this general statement is the Last-1 model, whose BIC scores are best for subjects with very low $\bar{\tau}$s. Again, these are the subjects who also have high BIC scores and whose switching behavior seems to be nearly independent of prior outcomes.

One might also hypothesize that the HH-5 model accurately reflects the behavior of subjects who may switch early in a session but then settle on an arm and stop switching all together. As Figure 7 shows, however, this does not appear to be the case. The figure sorts subjects by best-fitting model and then plots the length of each subject's last runs as a percentage of total trials. For each subject, the percentage is calculated by adding the lengths of the last runs of the three sessions and then dividing by the total number of trials in the session, $95+117+135=347$. (Note that the ES family of models is discussed below.)

Observe that average percentage was 42% for subjects for whom the Simple model fit best. In contrast, the average percentage was only 20% for subjects best described by HH-5. Thus, it does not appear that HH-5's BIC results reflect the fact that the model best captures the behavior of subjects who settle down early on an arm.

Exponential Smoothing. Last, Figure 8 compares the performance of ES to that of the Simple and Myopic-1 models.

The figure's left panel shows that, in the aggregate and for most of the range of the BIC scores, the ES-2 model outperforms the ES-1, the Simple, and the Mypoic-1 models. In the upper range of BIC scores—the cases in which model fit is poor—the ES-2 model

Figure 7 **Length of Last Runs as a Percentage of Total Trials, by Best-Fitting Model**

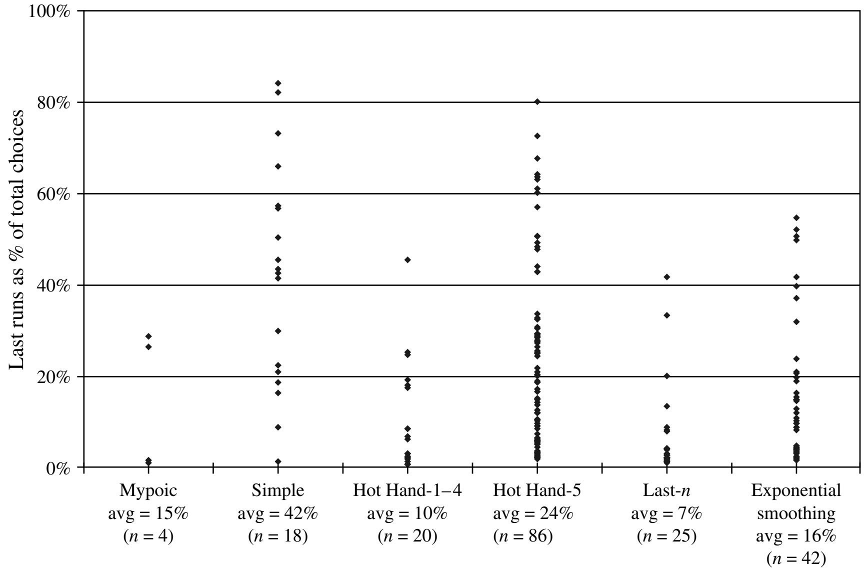
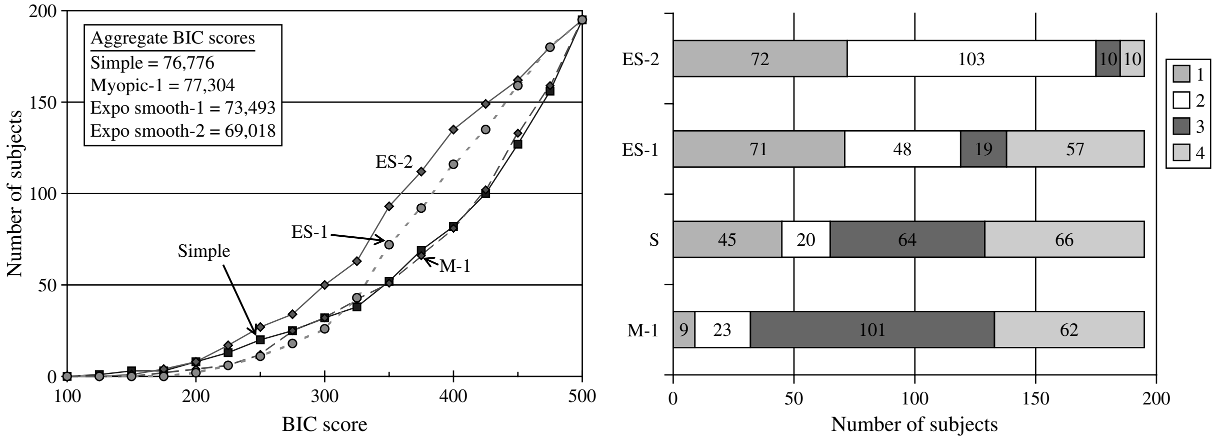


Figure 8 BIC Scores for ES Model



is tied with the ES-1 model. While the ES-1 model outperforms the Simple and Myopic-1 models, both in aggregate and in the above-325 range of BIC scores, all three models are dominated by ES-2.

The rankings in the right panel show that, on a subject-by-subject basis, the BIC scores of ES-2 dominate those of other models. Similarly, a one-sided paired t-test comparing ES-2 and ES-1 scores yielded a vanishingly small p-value, and an analogous Wilcoxon signed rank test returned a p-value of 0.01. Tests among other pairs of models—ES-2 versus Simple, ES-1 versus Simple, and ES-1 versus Myopic-1 returned p-values that ranged from highly significant (less than 0.001) to vanishingly small.

6.3.3. Summary of Results. Recall that we are interested in two dimensions of model performance. The first is fit to subjects' observed choices, as measured by both aggregate and individual-level comparisons of models' BIC scores. The second is tractability, the ability to derive simple (preferably closed-form) expressions for aggregate statistics—such as expected switching time or fraction of times chosen—from the models' primitive parameters. Table 1 summarizes our findings.

Among the Gittins index family of models, the Simple model performed best. While the BIC scores of Myopic-1 were on par with those of the Simple model, the Simple model is more tractable. The original Gittins index model of "rational" choice performed most poorly, both in terms of fit and tractability.

Among the descriptive models, the HH-5 performed best. Indeed, HH dominated both the Gittins index family and the other descriptive model in terms of fit, and it is also easy to analyze in terms of deriving useful aggregate statistics. The strong performance of the HH model must be qualified, however. As Figure 6 shows, the model's low BIC scores appear to be artifacts of longer run lengths.

There are other limitations to the HH model as well. While the HH's summary performance measures are easy to analyze in the context of Bernoulli outcomes, the model does not immediately apply to more complex reward distributions.[23] In contrast, the other models we have tested can be applied directly to generally distributed rewards. More fundamentally, HH-∞ is not applicable to analysis of supplier competition, because it models consumers as not responding to the quality of service they receive.

Of the second set of models that we tested, ES also performed quite well, both in terms of analytical tractability and model fit. ES compares well with the Simple model, with (roughly) the same level of analytical tractability and stronger BIC scores. We therefore favor the use of the ES model, and we elaborate on this recommendation below.

[23] One can generalize the HH model to handle rewards that are not Bernoulli distributed by considering "wins" to be outcomes above some threshold and "losses" to be outcomes below. This threshold would then become a free parameter for which to be searched.

Table 1 **Summary of Model's Fit to the Data and Tractability**

	Gittins index family				Descriptive models			
	G	M-0	M-1	S	Last-1	HH-5	ES-1	ES-2
Aggregate BIC	83,236	79,344	77,304	76,776	80,278	66,910	73,493	69,018
Tractability	−−	−	−	+	++	++	+	+

7. Discussion

Our experimental results are positive on two levels. First, they demonstrate that both the first and second difference of the average number of trials subjects spent on a given arm were increasing in the arms' expected reward. This finding is consistent with increasing, convex behavior predicted by the models of choice we have considered, as well as with more general industry observation (e.g., Jones and Sasser 1995). Second, they support the use of more analytically tractable models for use in the type of competitive analysis pursued in Gans (2002a), Gaur and Park (2003), and Hall and Porteus (2000). Both the ES and Simple models, in particular, are analytically tractable, flexible enough to be used with generally distributed rewards, and robust with respect to model fit.

We note that not all of the less elaborate models performed better in measures of fit. For example, the Last-n model fit the data poorly for higher n's, and the HH-n model performed poorly for lower n's. Thus, simplicity, by itself, does not appear to be a guarantee of a model's success.

The ES model particularly appeals to us for a number of reasons. First, it appears to fit the data well. Furthermore, the special case of $\gamma = 1$—which was excluded from consideration in our fit analysis—corresponds to a Last-1 model, so the model is (marginally) even more flexible that the results already indicate. Also, recall from §3.3 that ES is consistent with a model of a Bayesian subject that faces Markov-modulated rewards. We find this representation to be intuitively satisfying. Even though operations management papers often model the world as being stationary it is not; the ES models implicitly capture a subject's belief that his or her reward system is always changing. Still, both sets of results also raise important questions. In the following sections, we address two of the most visible concerns.

7.1. Tests for Run-Length Behavior

While our run-length results are consistent with predictions of convexity, there may be other explanations for the behavior that we observed. For example, the fact that the first difference between 0.65 and 0.40 is greater than that between 0.40 and 0.15 may be influenced by the fact that $\mathsf{P}\{\text{win}\} = 0.40$ for the competing arm and that subjects do not perceive gains equivalently to losses.

To rule out this hypothesis, one might also test cases in which the competing arm has a $\mathsf{P}\{\text{win}\}$ of 0.15 or 0.65, to see whether the increasing differences property still follows. One might also randomize treatment conditions differently, so that the runs of the first arms chosen could be compared on a within-subject basis.

More generally, it is worth emphasizing our first set of analyses only tests that observed behavior is consistent with the increasing-convex property. To *prove* that expected run lengths are increasing and convex in $\mathsf{P}\{\text{win}\}$, one would need to rule out all competing hypothesis. These types of tests await future research.

7.2. Tests of Model Fit

The fact that the HH-∞ model achieves the lowest BIC scores is, at first glance, discomfiting, and one may wonder if there is something about our experimental setup or analysis that should be changed to penalize this type of performance.[24] We believe it will be difficult to eliminate this effect. Because the BIC score is an adjusted measure of the likelihood of observing a given sample path, samples with low base rates of

[24] One may also wonder whether the Simple and ES models' free parameters reflect artifacts that are similar to the HH's. We have plotted both BIC score and average run length against free parameter values, and the results do not appear to reflect relationships that are more substantial than the original results for BIC scores and $\bar{\tau}$. For plots of BIC scores against $\bar{\tau}$, see Appendix H (online), and for plots of BIC scores against free parameter value, see Appendix I, (online).

switching tend to provide good fits for HH-∞. Conversely, HH-∞ should have a poor fit for subjects who switch frequently. Therefore, to properly penalize the HH-∞ model's behavior, one needs to set experimental conditions so that the observed switching rate among arms is higher.

How does one generate higher switching rates? The descriptive statistics reported in §4 suggest that more frequent switching is associated with lower win rates and small, rather than large, differences between the two arms' $\mathsf{P}\{\text{win}\}$s. In addition, we hypothesize that a smaller expected number of trials would increase switching, since a greater fraction of a subject's total winnings would depend on each trial.

Shorter sessions with high overall rates of switching would provide fewer, less discriminatory, trials over which the models' differences can be highlighted, however. That is, an experimental design with lower win rates and fewer trials is likely to push all models' BIC scores up toward that of random switching and make it more difficult for the various models' likelihoods to diverge. Therefore, we believe that a design including many trials with intermediate-level win rates that we used is a reasonable, if not perfect, experimental choice.

The use of a large number of trials has also allowed us to more clearly distinguish among the performance of models with various levels of learning. With many trials, models with more complete learning, such as the Gittins and Myopic models, tend to converge on an arm. Conversely, the ES, Last-n, and HH models represent a much weaker forms of learning, never converging on an arm. The Simple model represents something of a median, converging only in expectation, and only then if an arm meets the aspiration level μ^*.

The BIC scores suggest that subjects' choices are more consistent with models of choice that are both more myopic and attenuated in their form of learning. Aggregate measures of learning are also consistent with this finding. For example, on average, subjects switched 11.9 times in the first half of each session and 9.2 times in the second half. While this 22.4% decrease is important, it is far from reflecting a general convergence to a single arm.

While the Gittins index fared poorest in terms of BIC score, we recall that computational limitations have forced us to fix subjects' implicit (unobserved) discount rates to match the $\alpha = 0.99$ we used in the experiments. One may ask how much better the Gittins index would have performed had we had the ability search for the best-fitting implicit discount rate for each subject. A partial answer can be found by recalling that Myopic-0 is simply a Gittins index model with $\alpha = 0$. The fact that Myopic-0's BIC scores dominates the Gittins index's suggest that subjects *are* significantly more myopic than is optimal. Still, the questions concerning how much better the Gittins index would perform with an intermediate α remains open.

Section 6's results also suggest that there may be multiple segments of subjects, some of whose choices are more strongly guided by one model or another (see Figure 7). Furthermore, demographic or other, more easily observable data, may provide useful information on the type of model that best matches the pattern of her choice behavior.[25] For example, one significant difference we observed reflects subject gender: The 35% of women in our sample (69 out of 195) were much less likely than men to have their behavior best-fit by the Simple or Myopic rule(s).

Table 2 summarizes the data. A Fisher exact test in which counts for the Myopic and Simple model are included in a single "M or S" category yields a p-value of 0.006.[26] Thus, there does appear to be a potential gender difference in model fit, and it would be interesting to do follow-up work to better understand its nature and scope.

The existence of multiple segments, each best fit by a different model, is an important issue for future

[25] Similarly, for models with free parameters, such as Simple and ES, parameter values may also appear to be segmented.

[26] Fisher exact and χ^2 tests of category differences for the other demographic variables did not yield significant results at the 10% level. We also ran an OLS regression of earnings on the experiment's duration and $\bar{\tau}$ —controlling for demographics provided such as school, age, gender, ethnicity, and whether or not subjects reported they were confused by the experiment. The results are the same as those for a regression in which when we do not control for demographics: a unit increase in $\bar{\tau}$ is associated with an earnings increase of \$0.07; and a one-minute increase in the experiment's duration is associated with an earnings' increase of \$0.03.

Table 2 Numbers of Subjects for Whom Each Model Fit Best, as Measured by BIC Score

Gender	M or S	HH	Last-*n*	ES	Total
Female	1	42	9	17	69
Male	21	64	16	25	126
Total	22	106	25	42	195

experimental analysis, as well as for the application of our results. The fact that various segments of customers base their responses on different choice models affects overall market outcomes.

8. Conclusion

In this paper, we have examined the descriptive ability of Simple choice models in a bandit setting. We first developed a hierarchy of Bayesian models that range widely in complexity, from the Gittins index model that most closely represents rational behavior, through the simpler Myopic model, to the Simple model in which the customer's view of a supplier is representative of a category. We also considered other well-known choice models: The Last-*n* models, which are limited-memory analogues of the Myopic model; the HH models, which respond to sequences of losses on the current arm; and ES models, which take simple weighted averages of current and prior information.

We performed two sets of analysis, each intended to test the correspondence between subjects' behavior and the models' predictions. The first showed that average run lengths changed in manner that is consistent theoretical predictions: They should be increasing and convex in an arm's average reward. The second demonstrated that subjects' actual choices more closely matched the recommendations of less complex representations of choice.

Our results may also be of use in the larger task of building a positive theory of dynamic decision making, as described in Hutchinson and Meyer (1994). They suggest that there may be segments of subjects, some of whose choices are more strongly guided by one model of another. Similarly, a given subject may use more than one model of choice, perhaps shifting strategies over time. While we have not approached subjects' behavior in this fashion, the data exist that may make the analysis possible. Additional models, such as Q-learning, can be tested, as can more exploratory approaches. A recent paper by Houser et al. (2004) provides an exciting, new method for more exploratory analysis.

At the same time, there exist a number of limitations to our findings that bear repeating. While the experimental environment was essential for controlling the attributes of the arms and for generating predictions, as in any experiment, the artificiality of setting may lead to behavior that differs from the way people solve real world bandit-like problems. More narrowly, different experimental parameters, such as the expected number of trials ($(1-\alpha)^{-1}$), probabilities of success, or reward distributions may allow us to further discriminate among models' performance.

As in any experiment of this nature, there also exist more general natural limits to the external validity of our results. First, the bandit model, itself, may misspecify the basis of customer choice, though this (of course) was not a modeling problem these experiments were intended to address. For example, it may not be reasonable to assume that the quality distributions of the arms set by suppliers are independent. Second, our subject pool was drawn from the student population of a large university. While we have no reason to think that our subjects are nonrepresentative of students at this university, or that the behavior of students at the university is systematically different from that of the general population—as with any experiment that uses students as subjects—it is possible that our subjects chose arms differently than other populations would have.

Nevertheless, on the whole these results are positive for researchers in operations management. We originally viewed the models being tested as representing different trade-offs between analytical tractability and richness of the representation of the learning process. The experimental results suggest that subjects' learning is less strong than might be expected, however. It appears that the more tractable models that have been favored in operations management research can more than "adequately" capture the essentials of customer choice behavior.

Acknowledgments
Thanks to Gabriel Silvasi for developing the software for the experiments, as well as for providing computing support for the analysis of the data. Thanks also to Eliza Ponsz, who facilitated the testing of the software and recruiting of subjects. The authors are grateful to the anonymous referees and to Lawrence Brown, Dean Foster, Wes Hutchinson, and

Bob Meyer for suggestions concerning the analysis. They also thank seminar participants at the Yale School of Management and the University of Pennsylvania's Decision Processes Brown Bag Series for their helpful comments. The research of Rachel Croson was supported by NSF Grant SBR-9876079, which provided funds to pay the subjects who participated in the experiments. The research of Noah Gans was supported by NSF Grant SBR-9733739, The Wharton Financial Institutions Center, The Fishman-Davidson Center for Service and Operations Management, and the Wharton-INSEAD Alliance.

References

Allison, P. D. 1982. Discrete-time methods for the analysis of event histories. *Sociol. Methodology* **13** 61–98.

Anderson, C. M. 2001. Behavioral models of strategies in multi-armed bandit problems. Ph.D. dissertation, California Institute of Technology, Pasadena, CA.

Anderson, S. P., A. de Palma, J.-F. Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

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

Brezzi, M., T. L. Lai. 2002. Optimal learning and experimentation in bandit experiments. *J. Econom. Dynam. Control* **27** 87–108.

Brown, R. G., R. F. Meyer. 1961. The fundamental thoerem of exponential smoothing. *Oper. Res.* **9** 673–687.

Burns, B. D., B. Corpus. 2004. Randomness and inductions from streaks: "gambler's fallacy" versus "hot hand." *Psychonomic Bull. Rev.* **11** 179–184.

Bush, R. R., F. Mosteller. 1955. *Stochastic Models for Learning*. Wiley, New York.

Camerer, C. F., T.-H. Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* **67** 827–874.

Camerer, C. F., R. M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertainty* **19** 7–42.

Chang, F., T. L. Lai. 1987. Optimal stopping and dynamic allocation. *Adv. Appl. Probab.* **19** 829–853.

Erdem, T., M. P. Keane. 1996. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Sci.* **15** 1–20.

Erev, I., A. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique mixed-strategy equilibria. *Amer. Econom. Rev.* **88** 848–879.

Friedman, D., S. Sunder. 1994. *Experimental Methods: A Primer for Economists*. Cambridge University Press, Cambridge, UK.

Gans, N. 2002a. Customer loyalty and supplier quality competition. *Management Sci.* **48** 207–221.

Gans, N. 2002b. Service quality, switching costs and their effect on customer retention. Working paper, The Wharton School, OPIM Department, University of Pennsylvania, Philadelphia, PA.

Gaur, V., Y.-H. Park. 2003. Asymmetric consumer learning and inventory competition. *Management Sci.* Forthcoming.

Gilboa, I., A. Pazgal. 2001. Cumulative discrete choice. *Marketing Lett.* **12** 119–130.

Gilovich, T., R. Vallone, A. Tversky. 1985. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psych.* **17** 295–314.

Gittins, J. C. 1979. Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc.* **B41** 148–177.

Gittins, J. C. 1989. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, Chichester, UK.

Gittins, J. C., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani et al., eds. *Progress in Statistics*. North-Holland Publishing Company, Amsterdam, The Netherlands, 241–266.

Guadagni, P. M., J. D. C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing Sci.* **2** 203–238.

Hall, J., E. Porteus. 2000. Customer service competition in capacitated systems. *Manufacturing Service Oper. Management* **2** 144–165.

Harless, D. W., C. F. Camerer. 1994. The predictive utility of generalized expected utility theories. *Econometrica* **62**(6) 1251–1289.

Harrison, G. W. 1989. Theory and misbehavior of first-price auctions. *Amer. Econom. Rev.* **79** 749–762.

Henderson, P. W., R. A. Peterson. 1992. Mental accounting and categorization. *Organ. Behav. Human Decision Processes* **51** 92–117.

Horowitz, A. D. 1973. Experimental study of the two-armed bandit problem. Ph.D. dissertation, University of North Carolina, Chapel Hill, NC.

Houser, D., M. Keane, K. McCabe. 2004. Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica* **72** 781–822.

Hutchinson, J. W., R. J. Meyer. 1994. Dynamic decision making: Optimal policies and actual behavior in sequential choice problems. *Marketing Lett.* **5** 369–382.

Jones, T., E. W. Sasser, Jr. 1995. Why satisfied customers defect. *Harvard Bus. Rev.* **73**(6) 88–99.

Kahneman, D., A. Tversky. 1973. On the psychology of prediction. *Psych. Rev.* **80** 237–251.

Kass, R. E., A. E. Raftery. 1995. Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

Lehmann, E. L. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, CA.

March, J. G. 1996. Learning to be risk averse. *Psych. Rev.* **103** 309–319.

Matsuda, T., M. Sekiguchi. 1971. Models of human forecasting behavior: A note on the relationship between exponential smoothing and the Bayesian method. *J. Oper. Res. Soc. Japan* **13** 136–154.

Meyer, R. J., Y. Shi. 1995. Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Sci.* **41** 817–834.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity to process information. *Psych. Rev.* **63** 81–89.

Robbins, D., P. L. Warner. 1973. Individual organism probability matching with rats in a two-choice task. *Bull. Psychonomic Soc.* **2** 405–407.

Roth, A. E. 1988. Laboratory experimentation in economics: A methodological overview. *Econom. J.* **98** 974–1031.

Schmalensee, R. 1975. Alternative models of bandit selection. *J. Econom. Theory* **10** 333–342.

Simon, H. A. 1959. Theories of decision-making in economics and behavioral science. *Amer. Econom. Rev.* **49** 253–283.

Sutton, R. S., A. G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tversky, A., D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185** 1124–1131.