A Laboratory Study of Present Bias in Bandit Problems

Christopher M. Anderson*
Department of Environmental and Natural Resource Economics
University of Rhode Island
Kingston, RI 02881

September 20, 2000

Abstract

This paper uses a two-armed bandit to extend the result that people are present biased in search problems to the more general case of bandit problems. Bandit problems are economically significant, encompassing such phenomena as brand choice and natural resource exploration, and can help distinguish two models of present bias which are confounded in the search environment. Hyperbolic discounting attributes present bias to the discount sequence, and horizon truncation attributes present bias to a cognitive shortcut used to estimate the value of information gained through experimentation. Optimal behavior corresponds to placing a high initial value on experimentation, then decreasing it exponentially as information is gained in later periods. First period results from this study are consistent with present bias, suggesting that people do not experiment enough. Later period results suggest overexperimentation, a phenomenon which can be explained by horizon truncation, but is inconsistent with hyperbolic discounting. Together, these results suggest subjects place a moderate initial value on experimentation, then decrease it linearly. Implications for public policy and corporate strategy are considered.

1 Introduction

In many economically significant environments, agents must repeatedly choose among uncertain alternatives about which they can learn only through experimentation. Examples include the situations of a shopper deciding whether to

*This draft is preliminary. Comments welcome at cma@uri.edu. I am grateful for insightful discussions with and encouragement from Jeff Banks and Colin Camerer, and to the Russell Sage Foundation for financial support. I have also benefitted from helpful comments from an audience at the 1999 Meetings of the Southern Economic Association in New Orleans, and audiences at Caltech, Florida State University, Miami University and the University of Rhode Island. The instrument for the experiment presented here is available at http://eeps6.caltech.edu/˜cma/mab/UrnsMain.html.

purchase his favorite brand of orange juice or experiment with a new one he has never tried and an oil company deciding whether to continue testing a tract of land or to move its equipment to another tract. If these agents do not experiment enough, they can lose considerable welfare: the shopper could miss out on a delicious new brand of juice he would purchase and enjoy in the future, and the oil company may engage in an expensive recovery operation based on too few good test results. On the other hand, if these agents experiment too much, they may lose welfare as they pursue inferior choices.

Despite the economic importance of this sort of experimentation, little is known about how agents approach such problems. The existing knowledge is based on studies of search problems, which are a special case of experimentation problems. Agents in search problems do not search enough, suggesting that they are present biased—tempted to maximize their current payoff at the expense of future payoffs. However, the extant research leaves us without an understanding of how or whether present bias operates in the much broader domain of experimentation problems. This paper presents a laboratory experiment to determine if the present bias in search problems generalizes to the more sophisticated environment. If so, the experimental data can be used to test the predictions of two putative explanations of present bias, horizon truncation and hyperbolic discounting. Horizon truncation holds that agents think only a few periods ahead when making decisions, so present bias appears because the full future is not considered when calculating the future benefit of present experimentation. Hyperbolic discounting attributes present bias to a discount sequence which puts more weight on the current period than on future periods. Which of these models explains present bias has important practical implications for corporate strategy (e.g., natural resource exploration and new product marketing) and government policy (e.g., unemployment insurance and incentives).

1.1 Applications of the Experimentation Environment

Conceptually, experimentation problems focus on the value the agent assigns to the information obtained from experimentation. This information value arises from the expected increase in future payoffs based on the information. A surprising array of practical and economically significant decisions can be explained in terms of experimentation and information value:

Brand Choice: As mentioned above, a consumer shopping for a product he frequently buys, like orange juice or window cleaner, faces an experimentation problem: he must decide whether to purchase the best brand he's tried so far, or to experiment with new brands. He knows how good his favorite brand is on average, and how much it varies in quality, but he can learn about the new brand only by trying it. Therefore, he must consider whether the value of the information obtained about the quality of the new brand is worth foregoing his favorite brand. If he learns the new brand is better, he can use this information to improve his future utility by buying the new brand again. On the other hand, if it is worse, he has missed out on his favorite brand once, but he can return to it on the next purchase. If he underestimates the impact better orange juice will have on his future utility, he may never try the new brand and deprive himself of a possible gain.

Exploration: The oil company also faces an experimentation problem. Any agent exploring for natural resources tests land parcels to decide whether to mine or drill them. In this case, both additional testing and moving to a new tract are experimentation. The company can improve its estimate of how much oil is in the current tract with additional testing, or it can conclude that additional testing is so unlikely to influence its recovery decision that its equipment would be better used exploring another tract. If the company undervalues the information it would gain from additional testing on the current tract, it might decide to drill based on too few good test results, embarking on an expensive recovery operation in an area with few resources, or it might decide to abandon the parcel based on too few bad test results, leaving valuable resources in the ground.

Research and Development: Researchers want to allocate their time among a number of projects in a way that will maximize their chance of making an important discovery. For instance, a pharmaceutical company might experiment with several different approaches to treating a disease. The information acquired from experimentation can be used to focus subsequent research on the most promising alternatives, reducing the costs that they would incur by pursuing unpromising ones. However, if the company undervalues the information additional research on a specific treatment would provide, they may abandon an effective and profitable treatment whose promise was not immediately apparent.

Job and Price Search: Search problems are a special case of experimentation problems. Searcher's choices are somewhat different than those just described. Rather than repeatedly choosing from among multiple alternatives, at least one of which gives an uncertain payoff, searchers must decide whether to exit the problem with a known payoff stream (i.e., accept an offer) or to experiment by waiting for another offer. The information value here represents not the value of information *per se*, but rather the expected increase in future payoffs arising from the chance that future offers will be better.

For example, a worker looking for a job must decide to accept a wage offer, and receive that wage forever, or to experiment by continuing to look for a better offer. For low offers, she can expect to receive a better offer in the future, and this possibility constitutes the information value. If she does not experiment with enough different prospective employers, she could end up underemployed.

Similarly, a consumer looking for the best price on a product must decide whether to buy from the closest store at that store's price, or to experiment by searching other stores for a better price. The information value in this problem arises from the possibility that other stores have lower prices, and so the consumer may gain from searching. If the consumers do not experiment with different stores, stores can charge high prices, knowing consumers will not seek lower prices elsewhere.

1.2 Outline of this Paper

Determining if and understanding how present bias contributes to behavior in these environments is critical to helping agents maximize their welfare. The remainder of this paper is dedicated to establishing the role of present bias in experimentation problems. The next section presents evidence on present bias in search problems, and explains why those findings may generalize to experimentation problems. Section 3 discusses the two models of present bias to be considered here, hyperbolic discounting and horizon truncation. Section 4 formalizes the experimentation environment as a multi-armed bandit. Within the bandit framework, it presents a new theoretical result that the discounted present value of experimentation can be expressed as a number, the dynamic allocation index, even for the hyperbolic discounter. Section 5 describes the experiment, including the mechanism used to elicit dynamic allocation indexes from subjects. Section 6 reports the results of the experiment, using subjects' reported dynamic allocation indexes to draw conclusions about how present bias influences experimentation decisions. Section 7 discusses how the experimental results inform our understanding of economic experimentation and suggests directions for future work.

2 Evidence for Present Bias in Experimentation Problems

The possibility that present bias is a significant factor in experimentation problems should be of concern to economists because it implies considerable welfare is being lost because agents do not optimize. Suboptimal experimentation has already been observed in search problems. Cox and Oaxaca (1996, 1992, 1990, 1989) were concerned that job seekers may not engage in enough search and therefore end up underemployed. They asked experimental subjects to either accept a "wage offer" drawn from a known probability distribution, and receive that value in each remaining period, or take a fixed payment in the current period and receive another wage offer in the next period. They found that subjects' reservation wages were consistently lower than optimal, leading them to accept lower wages than optimal searchers would have taken. In this environment, the search cost is negligible, so Cox and Oaxaca attributed this early stopping to risk aversion.

This result replicated earlier studies by Schotter and Braunstein (1981) and Braunstein and Schotter (1982) which found that experimental subjects did not search enough. Schotter and Braunstein asked subjects to name a (nonbinding) reservation wage, and although their reservations wages were close to optimal, they spent significantly fewer than the optimal number of periods searching.

Although risk aversion could contribute to undersearch, there is evidence against its being the only explanation. First, Schotter and Braunstein induced a high level of risk aversion and still observed too little search. Second, in a less risky treatment where searchers were permitted to recall past offers, Cox and Oaxaca observed even less search, suggesting increased risk aversion. This replicated Hey's (1987) finding that reservation prices in an experimental price search were actually lower than without recall. This feature of the data is inconsistent with risk aversion since it implies the level of risk aversion varies within subjects across treatments. This, in turn, suggests some form of present bias may be contributing to reservation wage and price formulation in laboratory studies.

There is also some support for present bias in field studies of search. Although he did not consider a search-based model, the very high discount rates in appliance purchases observed by Hausman (1981) are consistent with under-search for quality. Similarly, Pratt, Wise and Zeckhauser (1979) observed that if consumers do not engage in enough price search, prices could vary widely from one retailer to the next. They measured the price variance of 39 goods selected at random from the Boston yellow pages by calling merchants selling each good. They found that price variance for moderate and high priced goods was in fact higher than could be sustained by optimal search, meaning people were paying supracompetitive prices for many goods. They attribute this suboptimality to an unobserved search cost. Present bias would make even a small search cost more salient, further reducing the amount of search.

However, Pratt, Wise and Zeckhauser's data also provide an indication that present bias may the result of some search and experimentation heuristic which does not perform well in the particular problems studied. They found people searched nearly as much for inexpensive goods like dry cleaning as for expensive goods such as boats, appearing to be more sensitive to the percentage that could be saved with search, rather than the monetary value of the savings. This suggests that the apparent present bias is not present bias *per se*, but rather an unintended consequence of a simple choice rule which is poorly calibrated to these problems.

Each of these results indicates that agents do not search enough, and that their search pattern is consistent with present bias. Because they disproportionately value the current period, present biased agents are more likely to stop searching and consume sooner, at the expense of future payoffs. Unfortunately, these results leave us with little information about whether present bias might operate in the broader domain of experimentation problems, including those in which there is more than one uncertain alternative.

Banks, Olson and Porter (1997) recognized the economic significance of experimentation problems and formulated a laboratory study to determine whether or not people behave optimally. They ran two treatments, one where myopic behavior, selecting the alternative with the highest expected value, was always optimal and another where it was sometimes optimal to choose the alternative with lower expected payoff. They observed a higher level of myopic behavior in the treatment where myopic behavior was optimal, suggesting a tendency toward optimality. However, they did not test for present bias explicitly, and a simulation study I have run suggests that their design is not powerful enough to distinguish optimal behavior from even very high levels of present bias.[1]

The search results can be easily extended to the more general experimentation framework. An experimenting agent must choose among several alternatives, at least one of which has uncertain outcomes. A searching agent must decide between accepting a fixed stream of payments (e.g., a price or wage offer) and experimenting with the distribution of offers. An agent tempted to maximize her current payoff is not induced to continue searching by the possibility of better offers. Similarly, the possibility that uncertain alternatives may pay better in the future is not enough to induce her to experiment with them; instead, she will opt for the alternative with the current highest expected payoff. In the earlier examples, this means the shopper will not try new brands of orange juice, the oil company will drill based on only a few good test results, and the pharmaceutical company will pursue only treatments which demonstrate early promise. In each case, these agents fail to maximize their future payoffs because they may miss delicious new brands of orange juice, signs that a tract will not be profitable or the true promise of a new treatment. One aim of this study is to test this intuition that present bias extends beyond search and causes welfare loss in experimentation problems.

3 Models of Present Bias

A second aim of this study is to identify a model which explains any present bias observed. I consider two models which are useful in different domains, hyperbolic discounting and horizon truncation. In hyperbolic discounting, present bias arises from a discount sequence which places relatively more weight on the present period than in standard exponential discounting. In horizon truncation, on the other hand, present bias arises from a cognitive shortcut in setting up and solving the dynamic programming problem whose solution yields the optimal strategy. It has been used to explain behavior in bargaining and dominance solvable games.

[1] The simulation study used a logit choice model to try to ascertain whether choice data could be used to draw conclusions about the β parameter in Equation 2. Then

$$Pr[y_t = 1] = \frac{e^{E[X|F_1]+\gamma(\lambda_1^* - E[X|F_1])}}{\sum_i e^{E[X|F_i]+\gamma(\lambda_i^* - E[X|F_i])}} \tag{1}$$

where $\lambda_i^* - E[X|F_i]$ is the optimal exponential information value for arm i. Using a variety of priors and payoff probabilities with Bernoulli arms, it was not possible to reject that $\gamma = 0$ and not reject that $\gamma = 1$ (the true model) in more than about 15% of the simulated datasets of 6000 choices; using a variety of priors and payoff distributions in normal arms, it was not possible to reject that $\gamma = 0$ and not reject that $\gamma = 1$ in more than about 40% of simulated datasets of 6000 choices.

The reason for this lack of power seems to be that, in practice, the difference between $E[X|F_i]$ and $E[X|F_{-i}]$ usually swamps the information value term, $\lambda_i^* - E[X|F_i]$. Very few choices between randomly valued arms are within the range of the information value to enable identification of the γ. From this, I conclude that choice data alone is not sufficient to draw conclusions about β; this is what motivates the somewhat complex design of the experiment presented in this paper.

3.1 Hyperbolic Discounting

Hyperbolic discounting attributes present bias to the discount function. Rather than behaving as exponential discounters, hyperbolic discounters have the time-separable utility function

$$\mathcal{U}(x_t, \ldots, x_T) = x_t + \beta \sum_{\tau=t+1}^{T} \delta^{\tau-t} x_\tau \quad \forall\, t. \tag{2}$$

A discount sequence of this form is also known as $\beta - \delta$ preferences.[2] Note that this discount sequence applies at every t, meaning there is an inconsistency between how the agent believes he will act in the future and how he actually does. Hyperbolic discounters believe they will be exponential beginning next period, but if $\beta < 1$, they place less weight on the value of future payoffs than would an exponential discounter. Given this discount sequence, it is assumed that they correctly set up and solve dynamic programming problems. This means that hyperbolic discounters will underestimate the value of experimentation because they heavily discount the future payoffs which benefit from present experimentation.

Hyperbolic discounting has been shown to explain a number of anomalous economic phenomena. Laibson (1997) shows that consumption and income fluctuate together because of hyperbolic discounting: people do not save enough to smooth their income because they are biased toward current consumption. He also shows that easier access to credit, and the concomitant possibility of increasing current consumption, led to declining savings rates during the 1980s. Additionally, O'Donoghue and Rabin (1999) show that Christmas clubs serve as "commitment devices" which help people resist the bias toward current consumption caused by the hyperbolic discount function; an exponential discounter, of course, would have no reason to pay a bank to prevent him from accessing his money until December.

Della Vigna and Paserman (1999) have taken a step toward extending these results to search and bandit problems. They used hyperbolic discounting to explain some aspects of field data on job search. They find that hyperbolic discounters do not want to incur a search cost, and so procrastinate their search efforts. Also, they reinforce the idea that Cox and Oaxaca's results could be due to hyperbolic discounting because once people do begin their search, hyperbolic discounters tend to take lower wage offers; they are more likely to end up underemployed because they stop their job search process too soon.

In each of these applications, there is a significant element of temptation: people are tempted to spend money they are holding rather than save it, and to take a job which begins paying now rather than continue searching. This temptation is often a significant factor motivating the application of the hyperbolic discounting model. In experimentation environments, there is no such

[2] This "quasi-hyperbolic" simplification of the hyperbolic discount sequence was introduced by Phelps and Pollak (1968). Lowenstein and Prelec (1993) discuss a more general hyperbolic discount function. See O'Donoghue and Rabin for a discussion of the differences between hyperbolic and quasi-hyperbolic preferences.

salient temptation. Therefore, discovering that hyperbolic discounting extends to experimentation problems would extend its domain considerably, and provide some evidence against the argument that such behavioral models are too problem specific.

A significant issue in hyperbolic discounting is how to handle the inconsistency between how an agent believes he will behave and how he actually does. For purposes of this study, I focus on the hyperbolic discounters whom O'Donoghue and Rabin call *naifs*. Naifs are "naive" about their own hyperbolic discounting tendencies and honestly believe that they will become exponential in the next period, although they do not; they are hyperbolic again.[3] Many argue that naifs should not remain naifs, that they should learn that they will be hyperbolic in the future. This objection has less bite in experimentation problems, where the cost of present bias may never be realized, especially if the agent never articulates to himself a commitment to be exponential in the future. For example, the hyperbolic shopper who bypasses the truly best orange juice every week in favor of the best brand he's had so far might never learn there is a better brand, and thus he would never regret his past purchases. Further, if he never promises himself he will try the new brand "next time," he may not realize that his eventual actions conflict with those he implicitly plans in computing an optimal strategy. Thus, experimentation problems are an important test for hyperbolic discounting because, unlike in the consumption and savings environment, even a potentially sophisticated hyperbolic discounter may never learn about his present bias.

3.2 Horizon Truncation

While hyperbolic discounting posits that present bias arises from the discount sequence, horizon truncation holds that present bias is a possibly unintentional side effect of a cognitive shortcut used to solve the dynamic programming problem. It says that, due to limited computational ability, laziness, or even a sophisticated cost-benefit analysis, agents do not consider the entire future when doing backward induction; rather, they perform a backward induction based on a short horizon, then add an adjustment factor to represent the value of omitted periods. If the adjustment factor is too small, horizon truncation leads to present-biased behavior because the agent considers only the value of experimentation represented in the abbreviated problem.

Horizon truncation appears in a number of domains. It is often employed deliberately in computer science to arrive at solutions to infinite horizon problems; if the future is discounted, computing several hundred periods into the future captures most of the value of a truly infinite horizon. In economic decision making, it has appeared in Rubinstein bargaining problems. Camerer et al. (1994) studied Rubinstein bargainers in an environment where the experimenters could observe which payoffs subjects considered when formulating their

[3] O'Donoghue and Rabin discuss various levels of hyperbolic discounters' self-awareness. The choice of naifs for this project is based on Laibson's results, but reinforced by the idea that bandits for self-aware hyperbolic discounters are intractable.

offers. Subgame perfection requires that subjects backward induct from the last stage payoff. Camerer et al. found, however, that subjects tend to look ahead only one stage, neglecting last stage payoffs entirely. These subjects were using a cognitive shortcut that required only the next stage's payoffs to formulate an offer.

Neelin et al. (1988) observed a similar phenomenon in alternating-offer bargaining. They looked at two, three and five period games. In the longer games, they observed the median first period offer was exactly the subgame perfect equilibrium of the two period game. In this case, subjects are using a two period truncated horizon, and not applying any adjustment for additional periods.

Behavior in dominance solvable games is also consistent with horizon truncation. In beauty contests (Nagel, 1995; Ho, Camerer and Weigelt, 1998), centipede games (McKelvey and Palfrey, 1992) and the dirty faces game (Weber, 1999), subjects obey only one to three levels of iterated dominance, which corresponds to a solving a truncated version of the game.

Applying the same cognitive shortcut to bandit problems could lead to present bias because the full future value of information acquired through experimentation is not represented. What is not clear, however, is how the adjustment factor responds to new information, the approach of the horizon, or the payoff scale. Improper sensitivity of this adjustment factor could explain bandit data which is not consistent with hyperbolic discounting. In addition, improper sensitivity to payoff scale could explain Pratt, Wise and Zeckhauser's observation that price search is insensitive to the amount to be saved.

One advantage of this paper's experimental approach is that it can distinguish hyperbolic discounting from horizon truncation, theories which are often confounded in field problems. If the environment is stationary, meaning the agent does not learn anything about the payoff distribution from receiving a draw from the distribution and the horizon does not approach, hyperbolic discounting and horizon truncation are not distinguishable. To a first approximation, job search and price search are both stationary, so these field studies could not distinguish the two models. The experiment presented here is designed to make a powerful distinction where these field studies cannot.

4 Formalizing the Experimentation Environment

To conduct a careful study of behavior in experimentation problems, the experimentation environment must be formalized. This section builds the theoretical foundations necessary to understand present bias in experimentation problems. First, it introduces the multi-armed bandit, a formal framework for studying experimentation. It then proceeds to explain how uncertain alternatives can be valued using a certain alternative: the expected payoff from a certain alternative which makes an agent indifferent between the certain and uncertain alternatives captures the discounted present value of present experimentation. A new theoretical result, that such a value exists for hyperbolic discounters, is presented.

4.1 Multi-armed Bandits

The experimentation problems described earlier can all be formally modeled as multi-armed bandits. The term bandit is used because each alternative can be thought of as a different slot machine. Each alternative, or arm, has two levels of randomness. First, an arm's payoffs are are randomly distributed. Second, one or more of the parameters of the arm's payoff distribution are unknown, but are drawn from known distributions themselves. In the case of the shopper looking for orange juice, his favorite brand which he has tried many times is a "known" average payoff arm, because he knows how much quality varies, and has a very clear idea of how good it is on average. The new brand, on the other hand, has unknown average payoff. The shopper has beliefs about how good it is on average, and about how much it varies, but he does not know for sure; he can update his beliefs by experimenting with the new brand.

In addition to a collection of arms, a multi-armed bandit must also have a discount sequence which indicates the present value of payoffs received in each future period. This is usually idiosyncratic to the agent. The agent combines her beliefs about the likelihood of different average payoffs with her beliefs about the variance of payoffs around the average to formulate a strategy which maximizes the present discounted value of payoffs received.

4.1.1 Information Value

The key concept in bandit problems, and the one which will eventually be used to identify present bias, is information value. The information value is the present discounted value of the expected increase in future payoffs arising from information gained by present experimentation. The consumer seeking orange juice can select the new brand, assuming its uncertainty, but also expect to gain from it. If the new juice is bad, he can switch back to his favorite brand next time. But if the new juice is good, he will have found a better juice, which he will buy and enjoy every period in the future and which he would not have found if he had not experimented. The information value captures the expected contribution to future payoffs arising from the possibility the new juice is better; it reflects the possibility the new juice is bad only in the present period because the shopper can switch back to his favorite brand.

If agents underestimate the information value, they will not experiment enough and may lock onto an alternative which gave good payoffs early, but which is not necessarily the one with the best average payoff. On the other hand, if agents overestimate the information value, they will experiment too much and waste choices on alternatives with low average payoffs. This intuition provides the basis for the experiment described in Section 5. It asks subjects for the information value they perceive from a single unknown arm. Their reported information value can be used to test for present bias by comparing it to the optimal information value for an exponential discounter.

4.2 Bandit Notation

For simplicity, attention is restricted to two-armed bandits. Otherwise, the notation I use largely follows that of Berry and Fristedt (1985) who were also concerned with variations in the discount sequence.

4.2.1 Arms

An arm consists of a distribution from which payoffs are drawn, a set of distributions from which the distribution of payoffs is selected and a prior over the set of distributions. Let $Q \in \mathcal{D}$ denote the distribution from which a payoff is drawn when the arm is chosen, where $\mathcal{D}$ is the set of possible payoff distributions. The agent's prior over the elements in $\mathcal{D}$ is denoted G. Although the theory given here works for general Q, $\mathcal{D}$ and G, those who prefer concreteness may consider Q to be a normal distribution with known variance σ^2 and unknown mean μ, $\mathcal{D}$ the set of normal distributions with variance σ^2 and $\mu \in \Re$, and G a normal distribution from which μ is drawn with known mean ν and known variance τ^2.

When an arm is selected, a payoff X is drawn from Q. The agent uses Bayes' rule to update her beliefs that Q is a particular element in $\mathcal{D}$. Let F on $\mathcal{D}$ denote the updated set of beliefs. Further, let $(X)F$ on $\mathcal{D}$ denote that the beliefs F have been updated to reflect the payoff X.

The two-armed bandits I consider will have one arm F, and a second arm with a known Q. Since Q will have only one parameter, the mean of the normal distribution, this known arm will be denoted λ, where λ is the value of the mean of the known Q.

4.2.2 Discount Sequences

A bandit consists of two elements: a collection of arms following the description above, and a discount sequence giving the discounted present value of payoffs in future periods. A general discount sequence will be denoted $A = (\alpha_1, \alpha_2, \alpha_3, \ldots)$, where α_t denotes the relative value of payoffs received in period t. In this notation, an exponential discount sequence is $A = (1, \delta, \delta^2, \ldots)$, and a hyperbolic discounter's discount sequence is $A = (1, \beta\delta, \beta\delta^2, \ldots)$. When it is convenient, $A^{(1)}$ will be used to denote the one-period-ahead continuation of A, $(\alpha_2, \alpha_3, \ldots)$.

Given these elements, the two-armed bandits on which this paper focuses can be written $(F, \lambda; A)$, where F is the unknown Q bandit, λ is the known Q bandit, and A is the discount sequence. Of particular interest will be the cases where A is exponential and hyperbolic, which will be denoted $(F, \lambda; \delta)$ and $(F, \lambda; \beta, \delta)$ respectively.

As mentioned above, this paper considers only naifs, hyperbolic discounters who honestly believe that they will be exponential next period, but then are not. The A notation for discount sequences does not adequately capture this, for it typically assumed that $A^{(1)} = (\alpha_2, \alpha_3, \ldots, \alpha_{T-1}, \alpha_T)$, but this is not the case for the naif. In fact, $A^{(1)}$ is A again, or if the horizon is finite, $A^{(1)} = (\alpha_1, \alpha_2, \ldots, \alpha_{T-1}, 0)$. This is not a problem for the analysis here because the naif acts on his (erroneous) belief in the present period; I only need to consider the problem he is solving.

4.2.3 Strategies, Worths and Values

A strategy in a bandit is a series of history-dependent arm selections σ, designating an arm choice in each period for each possible F in that period. The worth of a strategy (what it is expected to pay) is given by

$$W(F, \lambda; A; \sigma) = E_\sigma[\sum_{\tau=1}^{\infty} \alpha_\tau X_\tau] \tag{3}$$

where X_τ is the payoff received at time τ from whichever arm is prescribed by σ given the F at time τ.

The value of the bandit is the expected payoff given that the agent plays the optimal strategy (assuming it exists),

$$V(F, \lambda; A) = \sup_\sigma W(F, \lambda; A; \sigma). \tag{4}$$

Two other expressions of value are of interest. Let $V^F(F, \lambda; \beta\delta)$ be the value of selecting F in the current period and then continuing optimally and $V^\lambda(F, \lambda; \beta\delta)$ be the value of selecting λ initially and then continuing optimally.

$$V^F(F, \lambda; \beta\delta) = E[X|F] + \beta\delta E[V((X)F, \lambda; \delta)] \tag{5}$$

$$V^\lambda(F, \lambda; \beta\delta) = \lambda + \beta\delta V(F, \lambda; \delta) \tag{6}$$

These expressions will be useful in computing β.

4.3 Bandit Theory with Hyperbolic Discounting

The $(F, \lambda; A)$ bandit studied here was chosen because the λ arm can be used to value the F arm. The value of λ for which the agent is indifferent between selecting the two arms is what is known as a dynamic allocation index, or a Gittins index (Gittins, 1989), of arm F. The Gittins index is the sum of the expected payoff from F, $E[X|F]$, and an information value which reflects the expected gain to future payoffs arising from the information acquired through experimenting with F in the current period.[4]

For the consumer seeking orange juice, his longtime favorite brand would be "known" arm with "known" expected payoff λ. The new brand gives an uncertain payoff, so it is the F arm.

[4] The Gittins index is of particular interest in the case of exponential discounting and multiple uncertain arms. Gittins and Jones (1974) showed that if a Gittins index is calculated for each arm separately, the optimal strategy is to select the arm with this highest Gittins index in each period.

4.3.1 Hyperbolic Discounting and Optimal Stopping Problems

Actually solving bandits with a hyperbolic discount function is considerably more difficult than in the exponential case. The exponential problem can be (relatively) easily solved because it is an optimal stopping problem: once the agent chooses the λ arm, he will choose the λ arm in every remaining period (because nothing new is learned about F). This is not true for the hyperbolic discounter, however. She can choose the λ arm in the current period, believing she will experiment with the F arm in the next period. Without the optimal stopping property, solving the bandit is a far more (computationally) intensive process.[5]

Berry and Fristedt characterize the set of *regular* discount sequences, or those discount sequences for which a bandit is an optimal stopping problem. The following proposition confirms the intuition of the paragraph above that the hyperbolic discounter does not have an optimal stopping problem.

Proposition 1 *The hyperbolic discount sequence is not regular.*

Proof: Please see Section A.1.

Because regularity makes bandits tractable, most work has focused on regular discount sequences. Understanding how hyperbolic discounters should behave in bandits requires additional theoretical foundations.

4.3.2 Existence of an Optimal Strategy

First, it is important to know whether an optimal strategy exists. Berry and Fristedt use a standard argument to show that an optimal strategy exists for all possible discount sequences if there are a finite number of arms.

Theorem 1 *(Berry and Fristedt, 1985) There exists and optimal strategy σ^* for all possible priors G on $\mathcal{D}$ and all possible discount sequences A.*[6]

Their proof proceeds by demonstrating that there exists an optimal strategy for any finite horizon and then sending the horizon to infinity. For any finite horizon, there is a finite number of possible strategies (number of arms × length of horizon). Since any function has a maximum over a finite number of points, there is an optimal strategy for any finite horizon. Sending the length of the horizon to infinity gives general existence.

4.3.3 Existence of a Dynamic Allocation Index

The experiment described in Section 5 uses the dynamic allocation index, λ, as a measure of value for the F arm. In order for these inferences to be meaningful, it is necessary to establish that the dynamic allocation index represents the value of F for the hyperbolic discounter.

[5] Briefly, optimal stopping problems are simple because the continuation value of choosing the λ arm is $\sum_{\tau=0}^{T} \delta^{\tau} \lambda$, or $\frac{\lambda}{1-\delta}$ for infinite horizons. If the optimal stopping property does not hold, the value of choosing λ is a recursive calculation.

[6] This is a reader-friendly, if less precise, restatement of their Theorem 2.5.2.

Theorem 2 *For each nonincreasing discount sequence A with $A \neq 0$ and $\alpha_1 > \alpha_2$ and each distribution F on $\mathcal{D}$, there exists a unique function $\Lambda(F, A)$ such that the F arm is optimal initially in the $(F, \lambda; A)$ bandit if and only if $\lambda \leq \Lambda(F, A)$ and the λ arm is optimal initially if and only if $\lambda \geq \Lambda(F, A)$.*

Proof: Please see Section A.2.

This is the primary new theoretical result in this paper. The result based on the fact that $V(F, \lambda; A)$ is continuous and increasing in λ. This implies that $V^F - V^\lambda$ is strictly decreasing in λ. Roughly, this is true because λ is chosen earlier in the strategy sequence giving value V^λ. Because nothing is learned by choosing λ, the optimal sequence of λ choices giving $V((X)F, \lambda; A^{(1)})$ is similar to that giving $V(F, \lambda; A^{(1)})$. This proof is difficult because it is necessary to show that the value of information acquired from initial choice of F in V^F does not disrupt this relationship.

Given this result, the following proposition is easy to prove.

Proposition 2 *For a hyperbolic discounter with $\beta \leq 1$ and for each distribution F on $\mathcal{D}$, there exists a unique function $\Lambda(F, A)$ such that the F arm is optimal initially in the $(F, \lambda; A)$ bandit if and only if $\lambda \leq \Lambda(F, A)$ and the λ arm is optimal initially if and only if $\lambda \geq \Lambda(F, A)$.*

Proof: Please see Section A.2.

For $\beta < 1$, Theorem 2 establishes existence. For $\beta = 1$, the discount sequence is regular, so the existence result for regular discount sequences applies.

4.4 Properties of the Dynamic Allocation Index

Given that there exists a value of λ such that hyperbolic discounters are indifferent between F and λ, this value can be used to determine β in two ways. First, revealing that $\lambda = \ell$ makes them indifferent implies $V^F(F, \ell; \beta, \delta) = V^\lambda(F, \ell; \beta, \delta)$, where ℓ is the subject's reported dynamic allocation index. We can use Equations 5 and 6 to solve

$$\beta = \frac{\ell - E[X|F]}{\delta(E[V((X)F, \ell; \delta)] - V(F, \ell; \delta))}. \tag{7}$$

Because the values in the denominator are just stopping problems, their solution is not recursive. The expectation $E[X|F]$ is known, and ℓ is the value the agent reports as the dynamic allocation index.

Unfortunately, the quality of the approximation of the terms in the denominator is important, and accurate approximations are difficult if ℓ is substantially larger than λ^*, the optimal value of λ for the exponential discounter.[7] An alternative measure of β is the information value ratio. The information value ratio

[7] The reason is that if ℓ is large enough, then it is optimal to choose the λ arm initially in both the denominator terms unless the X in $E[V((X)F, \ell; \delta)]$ is very large; this low probability event determines the difference between the two terms in the denominator. Because the most common method of approximation is to truncate the distribution of payoffs near the tails, the error will be large relative to the values, meaning estimates of β will vary widely.

is

$$\mathcal{I}(F, \ell, \lambda^*) = \frac{\ell - E[X|F]}{\lambda^* - E[X|F]}. \tag{8}$$

This ratio does not give β, but it is always on the same side of one, so it is sufficient for present purposes. Information value ratios less than one suggest present bias, and information value ratios greater than one suggest a future bias.

5 Experimental Design

This experiment has two objectives. The first is to determine whether or not there is present bias in multi-armed bandits, and the second is to distinguish two possible causes of present bias. The existence of present bias can be established by comparing subjects' information values with the optimal information values of exponential discounters. This can be done by looking at the information value ratio, or by looking at β. Hyperbolic discounting requires that βs be constant as information is acquired and the horizon approaches, but horizon truncation, through its adjustment factor, allows for variations in β.

5.1 Incentive Compatible Dynamic Allocation Index Elicitation

Proposition 2 proves that there is a unique value of the known mean arm for which a subject is indifferent between the two arms. Equation 7 shows how a subject's ℓ can be used to determine β, which in turn can be used to test the predictions of hyperbolic discounting and horizon truncation. Thus the first design challenge of this experiment is to incentivize subjects to reveal truthfully the value of ℓ which makes them indifferent.

Proposition 2 claims that if $\Lambda(F, A)$ makes subjects indifferent, then they should pick the λ arm if its value is greater than $\Lambda(F, A)$, and F if λ is less than $\Lambda(F, A)$. One way to incentivize subjects' reported dynamic allocation indexes is to make choices for them based on their reported ℓs. For instance, if a subject reports $\ell < \Lambda(F, A)$ and the arm choice is based on ℓ, then there are values of the λ arm for which the λ arm would be chosen when the subject would prefer the F arm; if $\ell = \Lambda(F, A)$, there is no chance of this happening. This intuition suggests the following mechanism:

1. Endow each subject with an arm F with an unknown payoff distribution drawn from a set of distributions $\mathcal{D}$.
2. Explain to them that there is a second arm, λ, with a known average payoff of value λ which will be randomly drawn from some distribution with support $\Re$.
3. Before announcing the value of λ, ask each subject for a value ℓ_i, the minimum value of λ for which he or she would be willing to choose the λ arm in the current period.

4. Announce the value of λ. Fix the λ arm at that value for the remainder of the horizon.

5. For subjects with $\ell_i \leq \lambda$, force them to select the λ arm in the first period, but then allow them to proceed optimally, choosing the F and λ arms as they wish for all remaining periods. For subjects with $\ell_i > \lambda$, force them to select the F arm in the first period, but then allow them to proceed optimally, choosing the F and λ arms as they wish for all remaining periods.

Proposition 3 *Suppose $\Lambda(F, A)$, the dynamic allocation index for the arm F given A, exists and is unique. Then $\ell = \Lambda(F, A)$ is the unique optimal value of ℓ for a subject to report in the mechanism in this section.*

Proof: The proof follows the intuition given above and is presented in Section A.3.

Since Proposition 2 proves $\Lambda(F, A)$ exists for hyperbolic discounters, this mechanism can be used to elicit the dynamic allocation index in the first period of any bandit problem. However, once the value of the λ arm is known, the subjects need not report their true ℓ to receive the choice they want; this data would be much less reliable. A slight modification of the above mechanism can be used to get reliable ℓs in more than one period. Rather than revealing the value of λ in the first period, randomize the period in which the value of λ is revealed; subjects can be forced to pick F in the periods until λ is revealed. As long as the choice of ℓ affects the payoff with positive probability, subjects should still report ℓ truthfully. Because they do not have a choice if λ is not revealed, they cannot behave strategically. If λ is not revealed, subjects can use the payoff from F to update their beliefs about F and report a next period ℓ based on their updated beliefs. This allows collection of reliable data on a variety of beliefs, and with different horizons.

5.2 Bandits

This mechanism for truthfully eliciting dynamic allocation indexes requires a known mean arm λ and an unknown mean arm F. In this experiment, the F arm gives payoffs drawn from a normal distribution with $\sigma^2 = 100$ and a mean, μ, distributed $N(\nu, \tau^2)$ where $\nu = 1$ and $\tau^2 = 25$. The known mean arm also has variance of 100, to control for risk aversion.[8] Its mean is randomly selected from the same $N(1, 25)$ distribution as the mean of the unknown arm. The value is announced in the randomly determined period in which it is chosen.

Each bandit lasted for 10 periods, and each experimental session consisted of ten rounds. At the beginning of each round, new means for F and λ were

[8] These two levels of randomness in bandit arms have precise meanings in the terminology of risk and uncertainty. Risk is variance of the payoff distribution, and uncertainty is the variance in the distribution of the mean of the payoff distribution. These two arms are equally risky, so risk aversion cannot be a factor in behavior. What differs across arms is the level of uncertainty; uncertainty aversion may be a factor in this environment.

chosen. Subjects were told the shape and variance of the distribution from which their payoffs were drawn, as well as the shape, mean and variance of the distribution from which the mean of the payoff distribution was drawn. To emphasize the two-level nature of the randomness (i.e., that the mean of the distribution of payoffs itself has a distribution), the problem was posed as one of balls and urns, a familiar device for explaining randomness in experiments. Subjects were told there were two identical sets of urns with numbers on them; they could see the numbers on one set (the λs), but could not see the numbers on the other (the Fs). The payoff distribution was explained by saying there was an identical set of balls in each urn, and each ball had a number on it. The payoff was the sum of the number on the urn and the number on the ball. The probability distributions were conveyed using frequency tables, and by explicitly mentioning the parameters of the normal distribution in the instructions.

5.3 Other Design Features

Because I am primarily interested in how the information value behaves once subjects understand there is a value to experimentation, the instructions included a brief section about strategy.[9] Subjects were told that the information value arises from possible benefits in expected future payoffs, but were left to determine the magnitude on their own. To reinforce the instructions, the information value was featured on a quiz over the instructions, whose answers were explained before the experiment began, and during a guided practice period where the potential cost of an ℓ which is too low was emphasized.

To simplify the subject's task, and to make sure the difference between the reported ℓ and the expected value of F could be interpreted as an information value, subjects were provided with $E[X|F]$. The evidence that experimental subjects can effectively apply Bayes' rule is at best mixed (Kahneman and Tversky, 1972; see Camerer, 1995 for a review), so to avoid confounding my results with incorrect updating, I computed the Bayesian estimate of $E[X|F]$ and labeled it the "best guess" at the number on the unknown mean urn. Subjects were instructed that this "best guess" was arrived at using a law of probability called Bayes' rule.[10]

To encourage subjects to think carefully about the values of ℓ they reported, I used a bracketing mechanism to ask a sequence of questions to isolate the value of λ which made subjects indifferent between the two arms. I allowed values in [-15,30]. A test value, $\hat{\ell}$, was randomly chosen between these two endpoints. The subject was then asked, "Would you choose the [known mean arm] this period if its [known mean] were $\hat{\ell}$?" Subjects could click "Yes" or "No" buttons;

[9] A pilot run without this instruction suggested that it took a long time for subjects to realize there was an information value; including the instruction significantly reduced noise in the data. Whether or not people recognize that this value exists in general problems is a separate question.

[10] A few subjects explicitly rejected the best guess. Most claimed looking only at past payoff realizations provided a better estimate, suggesting that the neglect of base rates may be more than a cognitive shortcut.

"Yes" focused subsequent questions on lower values of $\hat{\ell}$, and "No" focused the search on higher values of $\hat{\ell}$. The questions continued with different values of $\hat{\ell}$, until the ℓ that made subjects indifferent was identified to the nearest 0.05 francs (0.4 cents).

To simplify analysis of the data, the random numbers used for payoffs were taken from a published random number table. This guaranteed randomness, but also ensured that each subject saw the same sequence of payoffs and arm values. This is important because, although computing the optimal index is a stopping problem, it is still computationally intensive. Having every subject make decisions based on the same set of beliefs greatly reduced the set of beliefs for which an optimal solution had to be computed.

5.4 Subjects

The subjects for this experiment were 23 Caltech undergraduates. Caltech undergraduates are a particularly good sample for this task because it is complex, and they have been selected for admission to Caltech because they are analytically gifted. They also represent a "best chance" for optimal strategies because they are more likely than other populations to be able to formulate and solve the dynamic programming problem which yields the optimal solution; if anyone does not need to use cognitive shortcuts, it is these subjects.

Payments to subjects averaged $20, with a maximum of $21 and a minimum of $10 for about 1 hour and 45 minutes of work. To verify that subjects understood the task, a debriefing questionnaire asked them to describe the task and their approach to it. Subjects' comprehension was good, except for two subjects who seemed to have difficulty with English and had to be excluded; these were also the two lowest-earning subjects. A third subject was excluded for answering "3" for almost every ℓ. Parts of the data from three other subjects were excluded. One subject said he was confused in the first four rounds and suggested his data be excluded. A second subject answered $\ell = 0.05$ for every query after the sixth round. A third subject expressed lexigraphical preferences, selecting $\ell \approx 30$ (the maximum allowed), and indicating on his debriefing questionnaire he would have selected higher had it been possible.[11] In each case, the data retained from these subjects are not idiosyncratic.

6 Results

Figure 1 shows the information values from a typical subject. Since the λ arm was introduced at random, each round provides a different amount of data: one period in rounds 1, 3, and 9, two periods in rounds 2 and 7, four periods in round 5, six periods in round 4, and seven periods in rounds 6, 8 and 10. Since each subject saw the same random number realizations, the ℓs elicited in each

[11] Interestingly, the minimum number of times he felt he needed to select F before considering λ decreased across rounds.

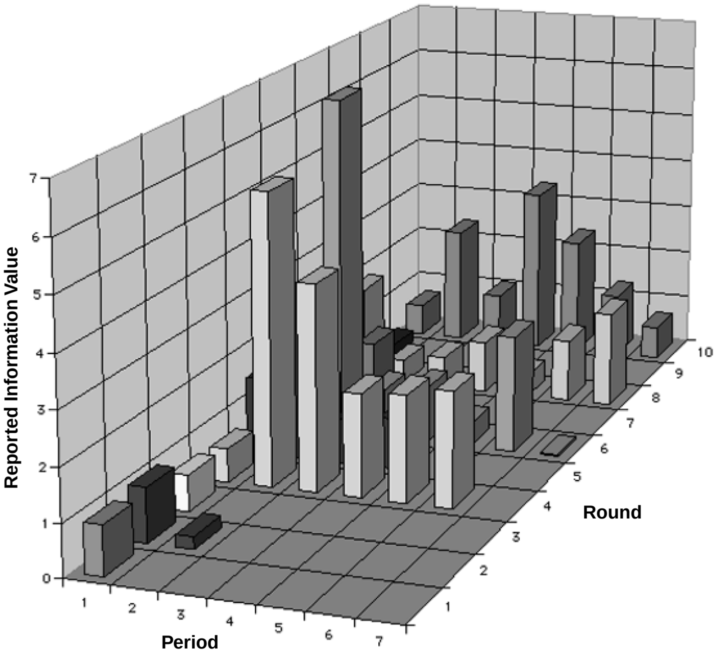

Figure 1: Information values reported by a typical subject.

round are based on the same payoff histories and thus can be aggregated or compared directly.

In optimal play, information values would begin at about 8.48 and then decrease, roughly exponentially, in later periods. The data do not follow this pattern. This subject's first period information values are low, around 1, a typical value for many subjects. This means the subject clearly understood that there was an information value, but did not have a good sense of its magnitude.

After the first period, this subject's information values fluctuate some, but generally decrease. This pattern was common. Subjects understood that the value of additional information fell as they learned more and the horizon approached, but they also tried to understand the effect of different information values. Testing different values was difficult because the F and λ arms were rarely close enough for a reasonable ℓ to indicate the wrong arm; this is not a flaw of the experimental design so much as a property of the bandit environment.

In this subject's data, rounds 4 and 6 are notable exceptions to the general pattern of decreasing information values. In these rounds, the true mean of the F arm was significantly negative, and this subject and many others were more hesitant than optimal to lower their ℓs in response to the expected payoff from the F arm.

Figure 2 presents a box-and-whiskers plot of the information value ratio defined in Equation 8, pooled by period across subjects and rounds. The box-and-whiskers plot indicates the distribution of the data at five points. The wide horizontal line indicates the median response in that period. The gray box covers the middle 50% of the data, and the "whiskers" cover the middle 90% of the data. The black dot in each period represents the mean response.

The overwhelming pattern in the data is that the information value ratios start below one, suggesting present bias, and increase as more information is acquired and the horizon approaches. At first glance, this is not consistent with hyperbolic discounting, which predicts that ratios should always be below one, and is consistent with horizon truncation with an adjustment factor which begins too small, and then does not adjust quickly enough.

This section's objective is to test which of the patterns in these pictures are statistically significant. If there is significant present bias, the data can be compared with the predictions of hyperbolic discounting and horizon truncation, giving insight into behavior in bandit problems.

Result 1 *First period ℓs are significantly below optimal, consistent with present bias.*

Support: Figure 3 shows the ℓs observed in the first period of each round.[12] This box-and-whiskers plot is interpreted the same as Figure 2, except that the whiskers cover only 80% of the data. The exponential-optimal value of 9.48 is indicated by the horizontal line spanning the graph. Only 25 of 182

[12] The subject with lexigraphical preferences is omitted from this graph. He chose a value at or near 30 every period and indicated that he would have chosen higher had it been possible to do so.

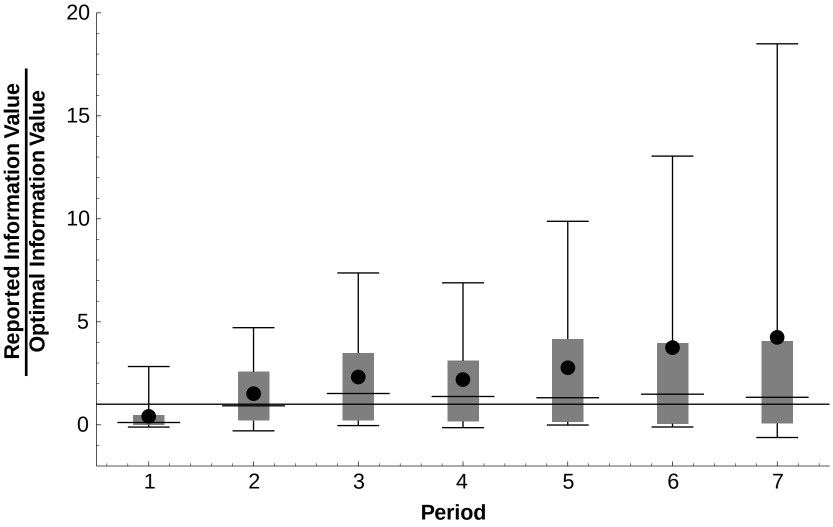

Figure 2: Box-and-whiskers plot of information value ratios across periods.

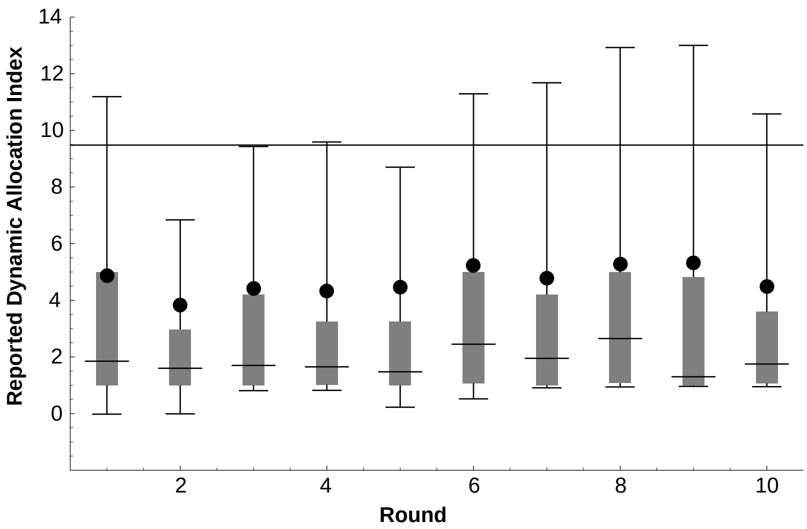

Figure 3: First period ℓs across rounds.

total observations are at or above the exponential optimum, and three subjects account for 20 of them. Based on this graph, it appears that first period ℓs are considerably below optimal.

That average choices are below optimal can also be tested on a subject-by-subject basis. Table 1 presents the means, standard errors and the p-values for the one-tailed t-test that mean information values are greater than or equal to the optimal value of 9.48. Even with this fairly small sample from each subject, the hypothesis that the mean is greater than or equal to the optimal value is rejected for 15 of the 19 subjects. This provides clear evidence for first period present bias.

To get some idea of what this level of present bias implies within the context of hyperbolic discounting, consider that the β that corresponds to an average response of 4.46 is 0.594. This is a little smaller than the $\beta = .70$ reported by Laibson (1997) in his field studies. However, it is not correct to interpret this as an average β because the transformation from ℓ to β is not affine; βs grow very quickly as the information value ratio exceeds one. The value of β at the mean ℓ is reported because it is very difficult to compute accurately the denominator of Equation 7 when the information value ratio is significantly above one; a couple outliers dramatically affect the mean.

Since this is an unfamiliar and somewhat abstract environment for subjects, it is possible that present-bias is an artifact of their unfamiliarity. If this is true, then they should learn to behave optimally, and thus appear less present biased, as they gain experience in the environment.

Result 2 *The first period information values do not increase in later rounds.*

Subject	Mean	Std. Err.	p-value
1	3.44	0.50	3.4E-07
2	1.13	0.02	3.4E-13
3	2.63	0.97	3.0E-05
4	9.00	0.37	0.11
5	8.30	3.22	0.36
6	1.93	0.17	4.8E-12
7	4.03	0.37	6.5E-08
8	5.00	0.03	5.5E-17
9	1.02	0.01	1.0E-24
10	-0.24	1.41	3.6E-05
11	2.56	0.20	4.2E-11
12	1.40	0.12	8.6E-14
13	0.33	0.21	6.1E-08
14	0.40	0.11	1.7E-14
15	11.59	2.85	0.76
16	-1.04	2.02	2.8E-04
17	3.04	1.01	6.4E-05
18	25.60	0.40	1.00
19	1.76	0.16	1.9E-12
Total	4.46	0.53	5.0E-18

Table 1: Subject-by-subject mean first period ℓs, with one-tailed t-test that $\mu_i \geq 9.48$.

Coefficient	Value	Std. Err.	t	$P > \|t\|$	95% CI	
γ_0	.481	.228	2.110	0.049	.002	.960
γ_{lag}	.848	.081	10.446	0.000	.678	1.019

Table 2: Results of lag regression. The summary statistics are F(1,18)=109.11 and $R^2 = .728$.

Period	Obs	Mean	95% CI		Median	95% CI	
1	182	0.41	0.29	0.53	0.11	0.08	0.17
2	131	1.50	1.20	1.81	0.92	0.75	1.22
3	95	2.31	1.66	2.96	1.56	0.80	2.16
4	97	2.21	1.64	2.78	1.37	0.77	2.28
5	78	2.77	1.94	3.60	1.45	0.71	2.26
6	78	3.74	1.96	5.51	1.58	0.85	2.05
7	58	4.25	1.90	6.61	1.36	0.27	2.03

Table 3: Mean and Median information value ratios for each period.

Support: To test whether first period information values approach optimality, I use a simple lag regression:

$$\mathcal{I}_t = \gamma_0 + \gamma_{lag}\mathcal{I}_{t-1} \quad \text{for} \quad t \geq 2. \tag{9}$$

Table 2 presents the results of this regression, with White-adjusted standard errors. If subjects were learning to increase their information values in the first period, γ_{lag} would be greater than one. The estimated γ_{lag} is not statistically greater than one; it is almost statistically *less* than one. The limit of this lag process is given by $\frac{\gamma_0}{1-\gamma_{lag}} = 3.16$, so only subjects with $\ell < 3.16$ were increasing their information values in later rounds; subjects with higher information values were decreasing them, on average. The one-tailed p-value for $\frac{\gamma_0}{1-\gamma_{lag}}$ being below the optimal value of 9.48 is 3.63×10^{-5}. Therefore, I conclude that subjects were not learning to increase their information values in later rounds, so first period present bias is robust to experience.

These results replicate the present bias observed in search problems, suggesting present bias affects bandit behavior. However, this experiment establishes some special circumstances which provide the opportunity to observe information values where they could not be observed in the field. Because subjects are forced to choose F when they would not have had there been another choice, we can learn about how information values change with beliefs and the horizon.

Result 3 *Second and later period mean information value ratios are higher than exponential optimal, suggesting a future bias, but median values are close to optimal. The shift from present bias to future bias cannot be explained by hyperbolic discounting.*

Support: Looking at the later periods in Figure 2, there seems to be a clear trend toward higher mean information value ratios as time passes. This intuition can be tested by looking at the mean responses in each period. Table 3 shows the mean information value ratios for each period. Every period after the first has a mean information value ratio significantly above one. Further, there is a clear trend toward higher ratios in higher periods; only between periods 3 and 4 is there a small (insignificant) decrease.

		Estimate	Std. Err	95% CI	
Segment 1	γ_0	0.57	0.86	-1.11	2.25
75%	γ_{belief}	-0.10	0.05	-0.20	0.00
	γ_{period}	0.32	0.28	-0.22	0.86
	α	-0.39	0.83	-2.02	1.24
Segment 2	γ_0	-0.35	1.66	-3.61	2.91
25%	γ_{belief}	-0.22	0.06	-0.34	-0.10
	γ_{period}	0.93	0.35	0.25	1.61
	α	0.66	0.42	-0.17	1.49

Table 4: Multicycle ECM estimates of two-segment regression model.

However, Figure 2 suggests the mean may not be the best description of the data. Although there is a clear upward trend in the mean, Table 3 indicates the medians are not statistically distinguishable from optimality (Mosteller and Rourke, 1973). This suggests some of the subjects have increasing information value ratios, but that most do not.

To test this two-segment population hypothesis, I use a multicycle expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993) to estimation a two-segment weighted least squares model on the second through seventh period information value ratios. The model regresses the information value ratio against the period number, controlling for $E[X|F]$. Heteroskedasticity is modeled by $\sigma_t^2 = \sigma^2 t^\alpha$, where α is a parameter to be estimated.

The objective is to find the two sets of model parameters and the assignment of subjects to parameter sets which is most likely given the data. My approach treats the parameter set which generates each subject's data as "missing data;" if I knew which subjects were in which segment, I could simply estimate the model separately on each segment. Instead, for any pair of parameter sets, the EM algorithm uses Bayes' rule to update an (estimated) prior to compute the relative likelihood that each parameter set generated each subject's choices. These probabilities are then used as weights to reestimate the two parameter sets. McLaughlan and Krishnan (1997) summarize the theoretical conditions under which iteratively updating probabilities and reestimating parameters converges to the maximum of the (complete data) log likelihood function.

Table 4 presents the parameter estimates for the two segments, as well as the size of each segment. As Figure 2 suggested, about a quarter of the population has significantly increasing information value ratios, while we cannot reject that information value ratios are constant for the other three quarters. For the first group, we can reject that hyperbolic discounting is the dominant factor in experimentation behavior. Because their information value ratios increase from below one to above one, we must conclude their βs do also, but this inconsistent with hyperbolic discounting. This test is not strong enough to reject hyperbolic discounting for the rest of the population, though the increase over the first period in their information value ratios suggests that their βs may be increasing

Period	Obs	Mean	Std. Err.	95% CI	
1	110	3.53	0.72	2.11	4.96
2	93	4.00	0.55	2.91	5.09
3	76	3.30	0.52	2.27	4.33
4	77	2.50	0.42	1.67	3.34
5	58	2.14	0.43	1.27	3.00
6	58	2.71	0.78	1.15	4.27
7	58	2.01	0.55	0.90	3.12

Table 5: Mean information values for each period for Rounds 5-10.

as well.

Horizon truncation, on the other hand, may not appreciate the extent to which the horizon approaches and may not fully appreciate the degree to which information acquired in the first period benefits later payoffs. The data are consistent with a model of horizon truncation with an adjustment factor which does not adjust enough as information is acquired and the horizon approaches. Hyperbolic discounting may still contribute to present bias, but only as the discount sequence of the truncated horizon problem.

One problem with the horizon truncation model as it is specified here is that it is not falsifiable. The model says very little about the "adjustment factor," and without restrictions on its possible values, any pattern of information value ratios is consistent with the model. One desirable feature in the adjustment factors is that they do not increase over time. A rough test of this, abstracting from the value computed for the shortened horizon, is that the information values decrease over time.

Result 4 *The information values decrease from period to period, consistent with an intuitive restriction on horizon truncation.*

Support: Table 5 presents the mean information values for Rounds 5 through 10. In these rounds, there is no significant increase in the mean information value from one period to the next. Including the first four rounds introduces a statistically significant increase from the first period to the second. Experience taught subjects with very low first period information values that they should be higher, and subjects who did not decrease their ℓs in response to negative $E[X|F]$s that they should be more responsive, so the difference is erased in later rounds.

Table 6 presents the results of the random effects regression on the last six rounds; these results are robust to the inclusion of the first four rounds. The significantly negative coefficient on period indicates that information values are declining across periods. Although this does not explicitly control for the information value computed from the truncated horizon, it does place an upper bound on the adjustment factor in each period. This is consistent with the restriction that the adjustment factor be decreasing from period to period.

Coefficient	Value	Std. Err.	t	$P > \|t\|$	95% CI	
γ_0	4.048	.792	5.111	0.000	2.496	5.601
γ_{belief}	-.152	.020	-7.525	0.000	-.192	-.113
γ_{period}	-.314	.091	-3.474	0.001	-.492	-.137

Table 6: Results of random effects regression of the information value on period for Rounds 5-10. The summary statistics are $\chi^2_2 = 66.0$ and $R^2 = .083$.

7 Discussion

This paper was designed to fill two gaps in our understanding of behavior in experimentation problems. First, it hoped to establish whether or not the present bias which has been observed in search problems is also represented in the more general environment. Second, given that present bias generalized, it hoped to distinguish between two competing explanations for present bias.

Looking at first period choices, the evidence from the experiment presented here supports present bias in the bandit environment. Later period evidence, however, suggests that agents do not remain present-biased as they acquire information; rather, most subjects behave nearly optimally, and a substantial portion of the population appears to become future biased. These results are consistent with observing only present bias in studies of search. The environments studied are stationary, so there is no opportunity to observe choices which, like later period choices in this experiment, reflect updated beliefs and an approaching horizon.

That agents rarely encounter such circumstances outside the lab may be a partial explanation for the later period overexperimentation observed in this experiment. Had the λ arm been available after the first period in every round, few subjects would have experimented in the second period. Few naturally occurring bandits force subjects to experiment. These results suggest that once he buys the new brand of orange juice, the shopper is more likely than optimal to buy it again. However, he is never forced to buy the new brand in the first place, and so never encounters his tendency to overexperiment.

This behavior poses a challenge for the policymaker, for she must decide whether it is worse to allow agents to continue to underexperiment, or to implement a policy which encourages initial experimentation, but which may lead to overexperimentation. I argue that, in almost every case, overexperimentation is the more desirable outcome. The reason is that the agent has both the incentive and the information to correct his behavior once he has experimented too much. Once he has bought the new orange juice a second time, he has the opportunity to regret his purchase and modify his behavior; he has both the incentive and information necessary to learn to experiment less. It is difficult to learn to experiment more, however, because the underexperimenting agent does not have the information necessary to determine that he is not optimizing; he does not know what he is missing.

Choosing the correct policy to help agents overcome their present bias in the first period requires understanding its cause. The second objective of this experiment was to test two models which predict present bias. A significant fraction of the population in the experiment made choice which are not consistent with hyperbolic discounting being the dominant factor in experimentation behavior. Hyperbolic discounting predicts that the value of β will be constant across periods. Although the information value ratio is not β, it is always on the same side of one. That the information value ratio is below one in the first period and above one in the later periods indicates β is also, meaning it is rising over time. This is inconsistent with the model, and we therefore must consider that hyperbolic discounting is not the only thing preventing agents from behaving optimally.

However, this does not mean that hyperbolic discounting is not a factor in experimentation problems outside the lab. Subjects in this experiment are paid one time at the end of the session, so time discounting *per se* seems an unlikely factor in behavior. Any discount rate effect would have to be attributable to some sort of time illusion in which their discount sequence is sensitive only to a number of periods, rather than the length of periods. Further, in some activities, such as saving for retirement, there are factors such as the advice of experts or cultural norms which influence people's behavior. These norms would not translate well to an abstract environment such as the one presented here.

Even if hyperbolic discounting is a significant factor in the field, its effect may be swamped by that of horizon truncation, which is the major cause of first period underexperimentation in this controlled environment. Rather than placing a high initial value on experimentation and then decreasing it exponentially as with optimal information values, subjects seem to place moderate initial value on experimentation and then decrease it linearly. This corresponds to a model of horizon truncation where the the adjustment factor is initially too small, but then does not decrease quickly enough as information is acquired. The data from this experiment are consistent with an intuitive restriction on this model, that the information values be decreasing over time.

It is important to note that these explanations augment Cox and Oaxaca's claim that risk aversion is the primary factor causing people to stop searching for jobs too soon. Because the two arms in this experiment are equally risky, risk aversion would predict optimal behavior in this experiment. That agents still do not experiment enough when faced with equally risky alternatives means risk aversion is not the only factor affecting search.

In some applications, whether hyperbolic discounting or horizon truncation causes present bias is critical in formulating public policy. Della Vigna and Paserman (1999) argue that job finding bonuses which reward the unemployed after they have held a job for a number of weeks may be too distant to significantly aid hyperbolic discounters. This is not necessarily true for horizon truncaters. Such an incentive would not influence their short-horizon value, but it may increase their "adjustment factor." Significant rewards in the future may influence the decisions of the present biased through the adjustment factor; before dismissing these programs as ineffective, this hypothesis should be investigated with field data.

In other applications, public policy is unnecessary because other agents in the economy may have an interest in helping agents overcome their present bias. For instance, companies introducing new brands and stores with low prices would both like consumers to experiment with them. If these agents know of consumers' present bias, they can take steps to encourage experimentation. A company with a new brand might offer free samples at the supermarket or through the mail, or generous coupons. Stores with low prices may aggressively advertise, or even, as some new dot-coms are doing, offer first purchases for free. These measures all encourage experimentation which will benefit the consumer in the long run.

If agents are aware of their present bias, they may also be willing to pay experts to help them avoid underexperimentation. While bandit problems are difficult to solve, an expert with a computer program can come much closer to optimality than this experiment has demonstrated even the most analytically capable non-experts can. Hiring an expert to make exploration decisions may significantly improve profitability by preventing costly overexperimentation or hasty recovery decisions.

In certain circumstances, individuals can also rely on expert advice. There is no shortage of expert advice on some intertemporal decisions, such as saving for retirement. Experts, both personal and in the media, constantly remind people to take advantage of tax incentives, employer matching plans. In this case, expert advice supplied by the private market and public policy are effectively combined so that people do not need to solve a dynamic programming problem. They can follow the experts' advice and will end up with an acceptable level of savings, if not one carefully tailored to their preferences.

Unfortunately, while some agents have incentives to assist present biased firms and consumers, there are also incentives to exploit them. Present bias suggests agents may be especially susceptible to bait-and-switch scams, or to misleading advertising. A consumer drawn to a store based on a low advertised price can easily be manipulated into buying a substitute product at a higher price because she is disinclined to conduct a price search on the new product. In these cases, public policy is needed to help present biased agents. Laws like the recent regulations requiring car dealers to clearly disclose down payment and financing information on leases can help present biased consumers avoid situations where their present bias would lead them to compromise their future welfare.

This experiment has demonstrated that agents do not experiment enough, and that their experimentation pattern is more consistent with horizon truncation than hyperbolic discounting. However, this is far from a complete picture of how present bias (and future bias) operate in experimentation environments. This experiment suggests a number of avenues for future research. First, this analysis does not indicate whether people are hyperbolic in their truncated horizon problem. Understanding if hyperbolic discounting plays a role in addition to horizon truncation requires more sophisticated econometric models and will be the focus of additional work in the near future.

Second, even this additional analysis will only detect if hyperbolic discounting is present in the laboratory environment, where there is no real time structure. If there is no time illusion, present bias could not be detected here, even if it is an economically important phenomenon. It would be nice to locate some field data which would allow a distinction between hyperbolic discounting and horizon truncation, but I suspect this would be difficult. A possibly easier avenue would be to design an experiment with true time structure, where decisions were made on a daily or weekly basis for a long period of time.

Third, these results suggest that agents formulate "adjustment factors" to account for periods they omit from any explicit solution to the problem. Pratt, Wise and Zeckhauser's observation that agents are not sensitive to the payoff scale of the search problem suggests that studying how the adjustment factors respond to changes in the environment may prove insightful. Additional laboratory work could systematically vary the variance of the arm payoffs, the variance of the priors, the payoff scale and the length of the horizon. Adjusting these factors could help focus policy efforts by identifying features of environments in which people are especially present biased.

Finally, future research might take an entirely different approach to the problem. This paper explores a complicated economic decision as a whole, abstracting from factors such as formulation of the dynamic programming problem, solving it through backward induction and Bayesian updating. Future work could use this experiment to direct inquiry into particular features of dynamic programming using simpler frameworks. This would allow study of problem components, which could then be built into a larger model of bandit behavior. If agents actually solve the problem by breaking it into these components, this could prove particularly valuable, and would have the added benefit of making discoveries which could be applied to a very wide range of problems.

With the additional information provided by these avenues of research, we can gain some insight into how people approach economic experimentation. This experiment confirmed the prediction that people are initially present biased; from an economic standpoint, there is little comfort in the fact that they later experiment too much because they do not experiment the first time to reach that stage. More work can expand this understanding, which can then be used to help craft policies and corporate strategies to aid individuals and firms in problems where the proper amount of experimentation is necessary to maximize welfare.

Citations

Banks, J., M. Olson and D. Porter. An Experimental Analysis of the Bandit Problem. *Economic Theory*, 10:55-77, 1997.

Berry, D. and B. Fristedt. *Bandit Problems.* New YorK: Chapman and Hall, 1985.

Braunstein, Y. and A. Schotter. Labor Market Search: An Experimental Study. *Economic Inquiry* 20:134-144. 1982.

Camerer, C. Individual Decision Making. In *The Handbook of Experimental Economics*, ed. A. Roth and J. Kagel. Princeton: Princeton University Press, 1995.

Camerer, C., E. Johnson, T. Rymon and S. Sen. Cognition and Framing in Sequential Bargaining for Gains and Losses. *Frontiers of Game Theory*, ed. K. Binmore, A. Kirman and P. Tani. Cambridge: MIT Press, 1994. 27-47.

Cox, J. and R. Oaxaca. Testing Job Search Models: The Laboratory Approach. In *Research in Labor Economics* vol 15. Greenwich, CT: JAI Press, 1996. 171-207.

——. Direct Tests of the Reservation Wage Property. *The Economic Journal*, 102:1423-1432, 1992.

——. Unemployment Insurance and Job Search. In *Research in Labor Economics* vol 11. Greenwich, CT: JAI Press, 1990. 223-240.

——. Laboratory Experiments with a Finite-Horizon Job-Search Model. *Journal of Risk and Uncertainty*, 2:301-329, 1989.

Della Vigna, S. and D. Paserman. *Job Search and Hyperbolic Discounting.* Mimeo. July, 1999.

Gittins, J. *Multi-Arm Bandit Allocation Indicies.* New York: John Wiley and Sons, 1989.

Gittins, J. and D. Jones. A Dynamic Allocation Index for the Sequential Design of Experiments. In *Progress in Statistics*, ed. J. Gani et al. Amsterdam: North Holland, 1974. 241-66.

Hausman, J. Individual Discount Rates and the Purchase of Energy-Using Durables. *Bell Journal of Economics* 10:33-54. 1979.

Hey, J. Still Searching. *Journal of Economic Behavior and Organization* 8:137-144. 1987.

Ho, T., C. Camerer and K. Weigelt. Iterated Dominance and Iterated Best Response in Experimental p-Beauty Contests. *American Economic Review* 88: 947-969. 1998.

Kahneman, D. and A. Tversky. On Prediction and Judgement. *ORI Research Monograph 12.* 1972.

Laibson, D. Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, 112:443-77, 1997.

McKelvey, R. and T. Palfrey. An Experimental Study of the Centipede Game. *Econometrica* 60:803-836. 1992.

McLaughlan, G. and T. Krishnan. *The EM Algorithm and Extensions.* New York: John Wiley and Sons, 1997.

Meng, X. and D. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* 80:267-278.

Mosteller, F. and R. Rourke. *Sturdy Statistics.* Reading, MA: Addison-Wesley Publishing, 1973.

Nagel, R. 'Unravelling in Guessing Games: An Experimental Study. *American Economic Review* 85:1313-1326. 1995.

Neelin, J., H. Sonnenschein and M. Spiegel. A Further Test of Noncooperative Bargaining Theory: Comment. *American Economic Review* 78: 824-836. 1988.

O'Donoghue, T. and M. Rabin. Doing it Now or Later. *American Economic Review* 89:103-24, 1999.

Phelps, E. and R. Pollak. On Second-best National Saving and Game-equilibrium Growth. *Economic Studies* 35:185-199, 1968.

Lowenstein, G. and D. Prelec. Preferences for Sequences of Outcomes. *Psychological Review.* 101(1): 91-108, 1993.

Pratt, J., D. Wise and R. Zeckhauser. Price Differences in Almost Competitive Markets.

Schotter, A. and Y. Braunstein. Economic Search: An Experimental Study. *Economic Inquiry* 19:1-25. 1981.

Weber, R. "Uncommon Knowledge: An Experimental Test of the Dirty Faces Game." Mimeo, 1999.

A Proofs of Propositions

A.1 Non-Regularity of the Quasi-hyperbolic Discount Function

Berry and Fristedt (1985) characterize the set of discount sequences for which a bandit reduces to an optimal stopping problem. Knowing this is important because optimal stopping problems are much better understood, and much easier to compute solutions for, than the general bandit problem.

Definition 1 *For any discount sequence $A = (\alpha_1, \alpha_2, \alpha_3, \ldots)$, let $\gamma_t = \sum_{\tau=t}^{\infty} \alpha_\tau$. Then A is* regular *if, for $t = 1, 2, \ldots$*

$$\frac{\gamma_{t+2}}{\gamma_{t+1}} \leq \frac{\gamma_{t+1}}{\gamma_t} \tag{10}$$

provided that $\gamma_{t+1} > 0$.

Unfortunately, intuition tells us the quasi-hyperbolic discount sequence may not be regular. The hyperbolic discounter is tempted to put off experimentation to next period, taking the known-mean arm now; while he selects the known-mean arm in the current period, he expects he will return to experimenting in the next period. The next proposition confirms this intuition.

Proposition 1 *The quasi-hyperbolic discount sequence is not regular.*

Proof: First I compute γ_1, γ_2 and γ_3, then I use these to check the definition of regularity. Note that the choice of $t = 1$ is important here, for choosing $t \neq 1$ does not contradict regularity; proving the definition is not satisfied only requires locating one t for which the condition is not satisfied.

From the definition of quasi-hyperbolic discounting, we have

$$\begin{aligned}
\gamma_1 &= 1 + \beta\delta + \beta\delta^2 + \ldots = 1 + \beta\delta \sum_{\tau=0}^{\infty} \delta^\tau = 1 + \frac{\beta\delta}{1-\delta} \\
\gamma_2 &= \beta\delta + \beta\delta^2 + \ldots = \beta\delta \sum_{\tau=0}^{\infty} \delta^\tau = \frac{\beta\delta}{1-\delta} \\
\gamma_3 &= \beta\delta^2 + \beta\delta^3 + \ldots = \beta\delta^2 \sum_{\tau=0}^{\infty} \delta^\tau = \frac{\beta\delta^2}{1-\delta}
\end{aligned}$$

Now plugging these into the definition of regular, we have

$$\begin{aligned}
\frac{\frac{\beta\delta^2}{1-\delta}}{\frac{\beta\delta}{1-\delta}} &\leq \frac{\frac{\beta\delta}{1-\delta}}{1 + \frac{\beta\delta}{1-\delta}} \\
\delta &\leq \frac{\beta\delta}{1-\delta+\beta\delta} \\
1 &\leq \frac{\beta}{1-\delta+\beta\delta} \\
1-\delta &\leq \beta(1-\delta) \\
1 &\leq \beta
\end{aligned} \tag{11}$$

Hence, the quasi-hyperbolic discount function is only regular if $\beta \geq 1$, which corresponds to the special case of exponential discounting; a quasi-hyperbolic discounter with $\beta < 1$ does not have a regular discount function. ♡

A.2 Existence of a Dynamic Allocation Index

This section proves Proposition 2. This is needed to show that there is a value of λ for which $V^F(F, \lambda; A) = V^\lambda(F, \lambda; A)$ for the hyperbolic discounter. There are several steps to this proof. First, I explain a result from Berry and Fristedt that $V(F, \lambda; A)$ is continuous and nondecreasing in λ. Then I prove an original result that $V^F(F, \lambda; A) - V^\lambda(F, \lambda; A)$ is nonincreasing in λ. This does most of the work in proving the proposition. I then show that if $\alpha_1 > \alpha_2$ then $V^F(F, \lambda; A) - V^\lambda(F, \lambda; A)$ is strictly decreasing in λ. Using this I show that there exists a value of λ for which $V^F(F, \lambda; A) = V^\lambda(F, \lambda; A)$ for any $\alpha_1 > \alpha_2$. The proposition is a direct consequence of this result.

The first step is to show that $V(F, \lambda; A)$ is monotonic in λ.

Theorem 3 *(Berry and Fristedt, 1985) For all F and A, $V(F, \lambda; A)$ is continuous and a nondecreasing function of λ.*

Berry and Fristedt provide an adequate proof of this theorem, so I shall only offer some intuition for its truth. An increase in λ can affect the value function in two ways: it increases the value of arm λ whenever it is chosen, and it expands the set of F over which the optimal strategy prescribes the λ arm to include those of higher expected value. Given this, an increase in λ could not result in a reduction of the value function because an increase in the value function never makes it more likely F will be chosen, and it strictly increases the value of any choice of the λ arm.

In order to show a dynamic allocation index exists, I also need a result about how the size of the error made by choosing the the wrong arm varies with λ. Define the function $\Delta(F, \lambda; A)$ as the difference in the value functions from choosing the F arm first and then continuing optimally and choosing the λ arm first and then continuing optimally;

$$\Delta(F, \lambda; A) = V^F(F, \lambda; A) - V^\lambda(F, \lambda; A). \tag{12}$$

The absolute value of the this quantity can be thought of as the cost of making an error by selecting the wrong arm initially. This quantity turns out to be very important, as the following lemma does most of the work in proving Proposition 2.

Lemma 1 *$\Delta(F, \lambda; A)$ is nonincreasing in λ when A is nonincreasing with $A \neq 0$.*

Proof: This proof is based on Berry and Fristedt's proof for Bernoulli F.

Fix $\lambda^* > \lambda$.

This proof proceeds in three parts. Part (i) derives an expression for $\Delta(F, \lambda^*; A) - \Delta(F, \lambda; A)$. Part (ii) performs a finite induction on the horizon to establish that $\Delta(F, \lambda^*; A) - \Delta(F, \lambda; A)$ is nonpositive. Part (iii) extends the result of Part (ii) to infinite horizons.

(i) The value of choosing the F arm first and then proceeding optimally is given by

$$V^F(F, \lambda; A) = \alpha_1 E[X|F] + E[V((X)F, \lambda; A^{(1)})]. \tag{13}$$

Similarly, the value of selecting the λ and then continuing optimally is given by

$$V^{\lambda}(F, \lambda; A) = \alpha_1 \lambda + V(F, \lambda; A^{(1)}). \tag{14}$$

Now define two more functions, which will prove to be of considerable algebraic convenience. $\Delta^{+}(F, \lambda; A) = \max[0, \Delta(F, \lambda; A)]$ and $\Delta^{-}(F, \lambda; A) = \max[0, -\Delta(F, \lambda; A)]$ so that

$$V^{F}(F, \lambda; A) = V(F, \lambda; A) - \Delta^{-}(F, \lambda; A) \tag{15}$$

$$V^{\lambda}(F, \lambda; A) = V(F, \lambda; A) - \Delta^{+}(F, \lambda; A). \tag{16}$$

Mnemonically, Δ^{-} is nonzero when $\Delta(F, \lambda; A)$ is negative, or when λ is the optimal arm.

Using these definitions, substitute for $V((X)F, \lambda; A)$ and $V(F, \lambda; A)$ in Equations 13 and 14 above. This gives

$$V^{F}(F, \lambda; A) = \alpha_1 E[X|F] + E[V^{\lambda}((X)F, \lambda; A^{(1)}) + \Delta^{+}((X)F, \lambda; A^{(1)})] \tag{17}$$

$$V^{\lambda}(F, \lambda; A) = \alpha_1 \lambda + V^{F}(F, \lambda; A^{(1)}) + \Delta^{-}(F, \lambda; A^{(1)}). \tag{18}$$

These expressions can then be used to compute $\Delta(F, \lambda; A)$. The first two terms in Equation 17 represent the value of selecting arm F in the first period, arm λ in the second and then continuing optimally. Similarly, the first two terms in Equation 18 represent the value of selecting arm λ in the first period, arm F in the second and then continuing optimally. Given this interpretation, subtracting the first two terms in Equation 18 from those in Equation 17 gives $(\alpha_1 - \alpha_2)[E[X|F] - \lambda]$. This gives

$$\Delta(F, \lambda; A) = (\alpha_1 - \alpha_2)[E[X|F] - \lambda] + E[\Delta^{+}((X)F, \lambda; A^{(1)})] - \Delta^{-}(F, \lambda; A^{(1)}). \tag{19}$$

Using this expression to compute $\Delta(F, \lambda^*; A) - \Delta(F, \lambda; A)$ gives

$$\begin{aligned} \Delta(F, \lambda^*; A) - \Delta(F, \lambda; A) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^{+}((X)F, \lambda^*; A^{(1)}) - \Delta^{+}((X)F, \lambda; A^{(1)})] + \\ \Delta^{-}(F, \lambda; A^{(1)}) - \Delta^{-}(F, \lambda^*; A^{(1)}). \end{aligned} \tag{20}$$

(ii) Proving the lemma requires that the difference in Equation 20 be nonpositive. This section performs induction on a finite horizon to demonstrate that this is true.

Let A_n be a nonincreasing discount sequence with finite horizon n, so elements after the n^{th} are zero.

First, suppose $n = 1$. Then, for all A_1, $\Delta(F, \lambda^*; A_1) - \Delta(F, \lambda; A_1)$ is nonpositive implies

$$\begin{aligned} E[X|F] - \lambda^* \leq E[X|F] - \lambda \\ \lambda^* \geq \lambda \end{aligned} \tag{21}$$

which is true by assumption.

Now suppose that the horizon $n > 1$ and $\Delta(F, \lambda^*; A_n) \leq \Delta(F, \lambda; A_n)$ for any nonincreasing A_n. Now I will use this induction hypothesis to show that $\Delta(F, \lambda^*; A_{n+1}) \leq \Delta(F, \lambda; A_{n+1})$.

Equation 20 can be rewritten with the truncated discount sequence

$$\begin{aligned}\Delta(F, \lambda^*; A_{n+1}) - \Delta(F, \lambda; A_{n+1}) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^+((X)F, \lambda^*; A_{n+1}^{(1)}) - \Delta^+((X)F, \lambda; A_{n+1}^{(1)})] + \\ \Delta^-(F, \lambda; A_{n+1}^{(1)}) - \Delta^-(F, \lambda^*; A_{n+1}^{(1)}).\end{aligned} \tag{22}$$

The first term on the righthand side of Equation 22 is nonpositive because $\lambda^* > \lambda$ by assumption and $\alpha_1 \geq \alpha_2$ by hypothesis.

The remaining two terms are nonpositive for similar reasons. Consider the second term. Since $A_{n+1}^{(1)}$ is nonincreasing and has horizon n, we have

$$\begin{aligned}E[\Delta^+((X)F, \lambda^*; A_{n+1}^{(1)}) - \Delta^+((X)F, \lambda; A_{n+1}^{(1)})] \\ = E[\max[0, \Delta((X)F, \lambda^*; A_n)] - \max[0, \Delta((X)F, \lambda; A_n)]].\end{aligned} \tag{23}$$

The induction hypothesis gives that $\Delta(F, \lambda^*; A_n) \leq \Delta(F, \lambda; A_n)$ for all F, in particular $(X)F$. Therefore, the second term in Equation 23 is always weakly larger than the first, implying that the second term in Equation 22 is nonpositive.

Consider the third term. Since $A_{n+1}^{(1)}$ is nonincreasing and has horizon n, we have

$$\begin{aligned}\Delta^-(F, \lambda; A_{n+1}^{(1)}) - \Delta^-(F, \lambda^*; A_{n+1}^{(1)}) \\ = \max[0, -\Delta(F, \lambda; A_n)] - \max[0, -\Delta(F, \lambda^*; A_n)].\end{aligned} \tag{24}$$

The induction hypothesis gives that $\Delta(F, \lambda^*; A_n) \leq \Delta(F, \lambda; A_n)$. Therefore, the second term in Equation 24 is always weakly larger than the first, implying that the third term in Equation 22 is nonpositive.

Since each of the three terms in Equation 22 is nonpositive, we conclude that the difference Equation 20 is nonpositive for every finite horizon. Now we let the horizon go to infinity to show it is nonpositive for infinite horizons.

(iii) Suppose $n = \infty$. Let A_T denote the truncation of A_∞ at finite T, so A_T coincides with A_∞ up to time T and has zeros afterwards. Letting $T \to \infty$ in the result from Part (ii) gives

$$\Delta(F, \lambda^*; A_\infty) \leq \Delta(F, \lambda; A_\infty). \tag{25}$$

Since $A = A_\infty$, we have $\Delta(F, \lambda^*; A) \leq \Delta(F, \lambda; A)$ for all horizons. This is sufficient to prove the lemma. ♡

Proving Proposition 2 requires a stronger version of Lemma 1.

Lemma 2 *If A is nonincreasing with $\alpha_1 > \alpha_2$, then $\Delta(F, \lambda; A)$ is strictly decreasing in λ.*

Proof: Parts (ii) and (iii) of the proof of Lemma 1 showed that each part of Equation 20 is nonpositive. If $\alpha_1 > \alpha_2$, then the first term on the righthand side of Equation 20 is strictly negative because $\lambda^* > \lambda$ by assumption. Therefore, Equation 20 is strictly negative and $\Delta(F, \lambda; A)$ is strictly decreasing in λ. ♡

Given this result, the existence of a dynamic allocation index is easy to prove. Proposition 2 follows immediately from the following theorem.

Theorem 2 *For each nonincreasing discount sequence A with $A \neq 0$ and $\alpha_1 > \alpha_2$ and each distribution F on $\mathcal{D}$, there exists a unique function $\Lambda(F, A)$ such that the F arm is optimal initially in the $(F, \lambda; A)$ bandit if and only if $\lambda \leq \Lambda(F, A)$ and the λ arm is optimal initially if and only if $\lambda \geq \Lambda(F, A)$.*

Proof: This proof begins by defining $\Lambda(F, A) = \inf\{\lambda \in \mathcal{D}$: the λ arm is optimal for the $(F, \lambda; A)$ bandit $\}$. Then I show that this definition implies that F is uniquely optimal if $\lambda < \Lambda(F, A)$. Then I use Lemma 2 to show that λ is uniquely optimal if $\lambda > \Lambda(F, A)$. Indifference at $\lambda = \Lambda(F, A)$ then follows from the continuity of V.

For $\lambda < \Lambda(F, A)$, we have

$$V^F(F, \lambda; A) > V^\lambda(F, \lambda; A) \tag{26}$$

from the definition of $\Lambda(F, A)$. Because $\Lambda(F, A)$ is the infimum value of λ for which λ is optimal, it must be that F is uniquely optimal.

The case where $\lambda > \Lambda(F, A)$ is a little harder because there may be values of λ above $\Lambda(F, A)$ where F is optimal. However, the fact that $\Delta(F, \lambda; A)$ is strictly decreasing in λ, as shown in Lemma 2, proves that this cannot be. Therefore

$$V^F(F, \lambda; A) < V^\lambda(F, \lambda; A) \tag{27}$$

for all $\lambda > \Lambda(F, A)$and λ is uniquely optimal.

Finally, if $\lambda = \Lambda(F, A)$, we have that neither F nor λ is uniquely optimal. Extending the continuity of $V(F, \lambda; A)$ to $V^\lambda(F, \lambda; A)$ and $V^F(F, \lambda; A)$, the previous cases sandwich possible values of $V^\lambda(F, \lambda; A)$ and $V^F(F, \lambda; A)$ to give

$$V^F(F, \Lambda(F, A); A) = V^\lambda(F, \Lambda(F, A); A), \tag{28}$$

which is equivalent to both arms being optimal initially for the $(F, \Lambda(F, A); A)$ bandit.♡

Proposition 2 *For a hyperbolic discounter with $\beta \leq 1$ and for A with $A \neq 0$ and each distribution F on $\mathcal{D}$, there exists a unique function $\Lambda(F, A)$ such that the F arm is optimal initially in the $(F, \lambda; A)$ bandit if and only if $\lambda \leq \Lambda(F, A)$ and the λ arm is optimal initially if and only if $\lambda \geq \Lambda(F, A)$.*

Proof: If $\beta < 1$, we have $\delta > \beta\delta$ for every δ. Therefore $\alpha_1 > \alpha_2$, so all the conditions of Theorem 2 are met.

If $\beta = 1$, Berry and Fristedt's Theorem 5.5.3 applies directly, providing an exact analog of Theorem 2 for regular discount sequences. Since then the discount sequence A is regular in this case, the conditions of their theorem are satisfied.$\heartsuit$

The existence of a dynamic allocation index should not be confused with the existence of an index result like that of Gittins and Jones (1976) which demonstrates that the optimal strategy is to select the arm with the highest index value. Indeed, it has been shown that this is not in general true for non-exponential regular discount sequences. I am not aware of any results, either positive or negative, for non-regular discount sequences.

A.3 Incentive Compatible Dynamic Allocation Index Elicitation

Proposition 3 *Suppose $\Lambda(F, A)$, the dynamic allocation index for the arm F given A, exists and is unique. Then $\ell = \Lambda(F, A)$ is the unique optimal value of ℓ for a subject to report in the mechanism in Section 5.1.*

Proof: This proof proceeds by showing that the mechanism of Section 5.1 induces an $(F, \lambda; A)$ bandit. Then I show that reporting an $\ell \neq \Lambda(F, A)$ lowers expected payoffs.

First, note that the mechanism of Section 5.1 provides for λ to remain the same for the rest of the horizon once it has been chosen. Therefore, an agent must maximize the payoffs from choices of either F or λ in each future period. Given that F and λ have the information structures of arms, and the agent has a discount sequence A, these elements form an $(F, \lambda; A)$ bandit. Therefore, we can use bandit theory, including that in Section 4, to assess the mechanism.

By definition of $\Lambda(F, A)$, we have $V^F(F, \lambda; A) = V^\lambda(F, \lambda; A)$ when $\lambda = \Lambda(F, A)$.

Suppose the subject picks $\ell = \Lambda(F, A) + \epsilon$ for some $\epsilon > 0$. Then suppose the random realization of $\lambda \in (\Lambda(F, A), \Lambda(F, A) + \epsilon)$ with positive probability, and suppose $\lambda = \Lambda(F, A) + \epsilon/2$ for specificity. Then, because $\ell > \lambda$, the subject must select arm F in period t. However, because $\lambda > \Lambda(F, A)$, λ is the unique optimal arm to play. This means $\Delta(F, \lambda; A)$ is negative, so $\ell = \Lambda(F, A) + \epsilon$ is not optimal. Uniqueness of $\Lambda(F, A)$ implies $\Delta(F, \lambda; A)$ is strictly decreasing in λ, so F is not optimal for any positive ϵ. Therefore, any value of $\ell > \Lambda(F, A)$ is not optimal.

The argument for $\ell < \Lambda(F, A)$ follows immediately, so the unique optimal value of ℓ is $\Lambda(F, A)$. Therefore, the mechanism induces subjects to truthfully reveal their dynamic allocation index.$\heartsuit$

B Instructions

You are about to participate in an experiment designed to provide insight into decision processes. The amount of money you make will depend partly on decisions you make and partly on chance. If you follow the instructions carefully and make good decisions, you might earn a considerable amount of money. You will be paid in cash.

B.1 How You Make Money

You make money by choosing an urn from which to receive a payoff. There are one billion hidden urns and one billion visible urns. Each urn has a number on its side. Each urn contains an identical set of one billion balls. Each of these balls has a number on it.

When you choose an urn, one ball will be randomly drawn from it. You will be told the *total* of the number on the ball and the number on the urn, but *not* the separate numbers. This total is your payoff, in francs.

B.2 Order of the Experiment

This experiment will proceed as a number of rounds. Each round will have exactly ten periods. At the beginning of each round, the computer will randomly select one hidden urn from which you can receive payoffs. You will not know the number on the hidden urn, but can learn about it by choosing the hidden urn. During a randomly determined period, one of the visible urns will be selected from which you can also receive payoffs. Unlike the hidden urn, you *can* see the number on the visible urn.

Even before a particular visible urn is selected, you must consider which of the values that could be on the visible urn would lead you to choose it. Each period, you will be asked for a cutoff value of the number on the visible urn, above which you would choose the visible urn and below which you would choose the hidden urn. This cutoff will be used to determine your choice in the randomly determined period in which one of the visible urns is selected: if the number on the selected visible urn is higher than your cutoff, the visible urn will automatically be chosen for you; if not, the hidden urn will automatically be chosen for you.

In each period, you must trade off choosing the visible urn, whose number you know, with learning more about the number on the hidden urn.

B.3 Urns

Other than being hidden, the set of one billion hidden urns is identical to the set of one billion visible urns. The Urn Number Table you have been given shows the number of urns with each possible number on it. The righthand column shows the percentage of each of the one billion urns with each possible number on it. For example, 10,203,858 urns, or 1.023on them.

Numbers are distributed among urns according to a bell curve, or in statistics, a normal distribution. The average of all the numbers is 1. The standard deviation is 10, meaning about 66(1+10) and -9 (1-10), and about 95

B.4 Balls

Each urn contains an identical set of one billion balls. The Ball Number Table you have been given shows the number of balls with each possible number on it. The righthand column shows the percentage of each of the one billion balls with each possible number on it. For example, 53,200,074 balls, or 5.32urn, have numbers between 4 and 5 on them.

Numbers are distributed among balls according to a bell curve, or in statistics, a normal distribution. The average of all the numbers is 0. The standard deviation is 5, meaning about 66(0+5) and -5 (0-5), and about 95

Note that the balls in each urn have several important properties:

1. Because the average number on the balls is 0, the average payoff you get from an urn is the number on the urn.

2. The distribution of balls is symmetric, which means the chance of getting one which increases your payoff by a certain amount is the same as getting one which lowers it a certain amount. For instance, the chance of an increase of 5 francs is the same as the chance of a decrease of 5 francs.

3. The chance of getting any particular ball is the same every period.

4. The chance of getting any particular ball is the same for each urn.

B.5 Visible Urn Cutoff

At the beginning of each period until one of the visible urns is selected, the computer will ask you "Would you choose the visible urn in this period if the number on it were [Number]?" If you would, click the "Yes" button, if not, click the "No" button. You will be asked a series of these questions, with a different [Number] each time, until the cutoff point at which you would just prefer the visible urn has been narrowed down to the nearest 0.05.

You should answer these questions carefully because, in the period in which a visible urn is selected, your urn choice will be made for you based on your answers. The computer assumes you will choose the visible urn for all numbers larger than the cutoff, and the hidden urn otherwise. Therefore, it will automatically choose the visible urn if the number on it is larger than the cutoff, and the hidden urn if the number on the visible urn is smaller than the cutoff.

B.6 Using the Computer

There are four panels on the computer screen. You may click in these panels with your mouse, but please do not attempt to use any other applications, look at the source code for this experiment or visit any other web sites during the experiment.

B.6.1 The History Panel

The long vertical panel on the left will contain your playing history. Please look at that panel now. For each period, it will show your choice of urn, your payoff and the visible urn cutoff; recent periods will be added to the top of the list, though earlier periods will still be accessible by scrolling down.

B.6.2 The Information Panel

Please look at the top of the three panels on the right side. It provides you with information on the current period, your total payoff and the number on the visible urn, if it has been selected. It also shows a *best guess* at the number on the hidden urn. The computer uses a law of probability, Bayes' Rule, to integrate the information in the urn number table and the ball number table with the payoffs you have received from the hidden urn to formulate a best guess at the number on the hidden urn. This number will change as you select the hidden urn and get more information about it.

B.6.3 The Urn Choice Panel

Please look at the middle of the three righthand panels (which now has a "Begin" button). This is where you indicate your choice of urn each period. To indicate your choice of an urn, click once with the mouse in the circle in front of the name of the urn you wish to choose; a black dot will appear within the white circle. Then click the *Submit* button at the bottom of the panel one time with the mouse. Clicking the Submit button causes the computer to select a ball and calculate your payoff for the period.

B.6.4 The Instructions Panel

The bottom of the three right panels will contain these instructions. You may scroll through them and examine them at any point during the experiment.

B.7 Summary

1. The experimenter will announce the beginning of the period.
2. If one of the visible urns has not yet been selected:
   (a) You will be asked a series of questions to determine the visible urn cutoff, the smallest number on the visible urn for which you would choose it that period.
   (b) There is a 3/10 chance the visible urn will be selected that period. If it is, the computer will automatically choose the visible urn for you if the actual number on the selected visible urn is larger than the cutoff, and the hidden urn if the actual number on the visible urn is smaller than the cutoff.

   If no visible urn is selected, you must choose the hidden urn.

If a visible urn has been selected, you can choose either the visible urn or the hidden urn.

3. A ball will be drawn from your chosen urn.
4. The number on the ball will be added to the number on the urn you chose to determine your payoff.
5. The computer will notify you of your payoff and update your history.
6. Record your choice and payoff on your Record of Earnings Sheet.
7. Wait for the experimenter to announce the beginning of the next period.

Francs will be worth $0.08 (8 cents) each. Feel free to earn as much money as you can. Are there questions?

B.8 Strategy

You want to allocate your ten selections among the two urns to maximize your total payoff. Since each urn has the same set of balls in it, if you knew the number on both urns, you would select the one with the higher number in each period.

Since you do not know the number on the hidden urn, it is helpful to learn about it from experience. If you choose the hidden urn several times, you get a pretty good estimate of its number. Choosing the visible urn, on the other hand, only gets you a payoff. You do not learn anything about the number on the hidden urn.

Given this, you should never select the visible urn if you think its number is lower than that of the hidden urn. However, you may want to choose the hidden urn even if the visible urn's number is higher than your best guess at the number on the hidden urn, especially if your beliefs about the hidden urn are based on only a couple of tries. Your belief that the hidden urn does not pay well may be the result of a couple bad balls, and more attempts may reveal it in fact pays better on average.

If you select the hidden urn a couple more times and it does not pay well, then you can switch to the visible urn. But if it turns out to pay well, then you will have found a way to get high payoffs which you would not have known about had you not chosen the hidden urn those few periods. Of course, it is possible that the visible urn will be enough better that the potential cost of trying the hidden urn is unlikely to be repaid with higher payoffs in the future. Exactly how good the visible urn has to be is your cutoff value, with the difference between the cutoff and your best guess representing the value of the information you get from choosing the hidden urn. How much you value the information depends on your beliefs about the number on the hidden urn, how much your beliefs are likely to change with one more attempt and the number of periods left to exploit what you have learned.

Thus, each period, you must trade off maximizing that period's payoff (by choosing the urn you currently believe to have the higher number) with refining your beliefs about the number on the hidden urn, impacting your future decisions and payoffs.