Behavioral Models of Strategies in Multi-Armed Bandit Problems

Thesis by

Christopher Madden Anderson

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2001

(Submitted September 25, 2000)

Acknowledgements

I have received help and support from many people during this research.

I would like to thank my committee, Colin Camerer, Jeff Banks, Dave Grether and Paolo Ghirardato for their support and thoughtful comments. I would especially like to thank Colin, my thesis advisor , and Jeff, for not just helping me with the theory presented here, but for teaching me how to develop it myself.

I have received helpful comments on parts of this thesis from many people and audiences. For Chapter 2, I received comments from Teck Ho, Richard McKelvey, Tom Palfrey and an audience at Caltech. For Chapter 3, I have received comments from audiences at Caltech, Florida State University, Miami University (Ohio), University of Rhode Island and the 1999 Meetings of the Southern Economic Association. For Chapter 4, I have received comments from the 2000 Fall Meetings of the Economic Science Association.

Colin Camerer and the Russell Sage Foundation have provided financial support for the experiments reported here.

Abstract

In multi-armed bandit problems, agents must repeatedly choose among uncertain alternatives whose true values they can learn about only through experimentation. Information acquired from experimentation is valuable because it tells the agent whether to select a particular option again in the future. Economically significant applications include brand choice, natural resource exploration, research and development and, as special cases, job and price search.

Despite the importance of these applications, little is known about whether firms and individuals appreciate the value of information in bandit problems. That which is known is based on laboratory and field studies of search problems. These studies suggest that people do not search enough, perhaps because of search cost or risk aversion. This thesis attempts to ascertain whether this undervaluation of information extends to the more general bandit environment, and, if so, whether the suboptimality is attributable to search cost, risk aversion, or some other cause.

The results of three laboratory experiments, each addressing a separate family of putative explanations for undervaluation of information in bandits, are presented. The first asks subjects to choose among a set of uncertain alternatives, controlling for mean-conditional risk and search cost. Although subjects appreciate that there is value to information, they experiment less than the optimal amount. Since there is no experimentation cost and mean-conditional risk is constant, these explanations cannot be the primary cause of underexperimentation.

The second experiment uses a more powerful design, asking subjects to report their Gittins indexes, rather than just make a choice. This additional information is used to test that agents are hyperbolic discounters who do not experiment enough because they are disproportionately tempted to maximize their current payoff at the expense of future payoffs. This, too, does not appear to be a primary explanation for underexperimentation because the agent's level of present bias changes over time,

contrary to an assumption of the model.

The third experiment tests whether ambiguity aversion, or distaste for variance in the distribution from which the means of the payoff distributions are drawn, contributes to undervaluation of information. Consistent with a prediction of ambiguity aversion, subjects have both lower-than-optimal Gittins indexes and higher-than-optimal willingness to pay for information about the true values of ambiguous alternatives. These results are not consistent with hyperbolic discounting, risk aversion or quantal response behavior. However, the errors vary only with changes in the bandit's horizon, not with small changes in mean and variance as ambiguity aversion predicts.

Taken together, these experiments suggest ambiguity aversion is a likely cause of suboptimal play in bandits, as is cognitive shortcuts used in formulating and solving the dynamic programming problem. If these errors can be demonstrated across a wide enough set of bandits, in the field as well as in the laboratory, then policies can be developed based on this behavioral understanding of choice. These policies can improve the welfare of the workers, shoppers and firms who have to solve bandit problems.

Contents

Chapter 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 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 do not appreciate the value of the information they gain from experimentation. However, the extant research leaves us without an understanding of how or whether this undervaluation operates in the much broader domain of experimentation problems.

This thesis presents a series of laboratory experiments to determine if the undervaluation of information in search problems generalizes to the more sophisticated environment. If so, the experimental data can be used to test the predictions of three putative explanations for suboptimality: horizon truncation, hyperbolic discounting and ambiguity aversion. Horizon truncation holds that agents think only a few periods ahead when making decisions, so undervaluation occurs because the full future is not considered when calculating the future benefit of present experimentation. Hyperbolic discounting attributes undervaluation to present bias—agents are tempted to maximize their current payoff at the expense of future payoffs because their discount

sequence puts more weight on the current period than on future periods. Ambiguity aversion attributes undervalution to a distaste for the variance in the distribution of the unknown mean of the payoff distribution. 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 Evidence for Suboptimal Information Acquisition in Experimentation Problems

Previous work on experimentation problems has established that agents do recognize that there is value to collecting information, but has not addressed the question of whether or not they place enough value on collecting information. Evidence from field and laboratory studies of search problems suggests that agents do not search enough, and therefore may not collect information in the more general experimentation environment.

1.2.1 Study of Experimentation Problems

The only study to address whether agents appreciate the value of information in bandits is Banks, Olson and Porter (1997). They 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. This means that agents do recognize the value in the information they obtain through experimentation. However, Banks, Olson and Porter do not consider whether or not the magnitude of the intertreatment shift they observed was consistent with optimal behavior.

1.2.2 Evidence from Search Problems

Although Banks, Olson and Porter's experiment does not address whether people appreciate the magnitude of the value of the information they gather through experimentation, studies of search problems suggest that agents may be losing welfare because they do not acquire enough information.

The possibility that agents underexperiment 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.

Cox and Oaxaca's baseline design asked subjects to engage in a 20-period job search from a known distribution of offers, uniform on integers 1 to 10. A search period consisted of a random draw to determine if the subject would get a wage offer, and if so, the draw of the offer itself. Once a subject accepted an offer, she was paid that wage in every remaining period. In different treatments, Cox and Oaxaca manipulated the length of the horizon, the probability of receiving an offer (1989), the size of the payment the subject received if no offer was recieved or an offer was rejected (1989,1990), and the variance of the offer distribution (1989). In other papers, they elicited a binding reservation wage (1992), and allowed subjects to recall previous offers, reducing the risk involved in continued search (1996).

Across these treatments, they found that subjects terminated their search sooner and with lower offers than the optimal model predicts, and that their reservation wages were lower than optimal.

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 reservation wages were close to optimal, they spent significantly fewer than the optimal number of periods searching.

In known offer distribution experiments, an analog to the information value arises

because future offers may be higher than the mean of the offer distribution. In another experiment, Cox and Oaxaca (2000) examine the effect of unknown offer distributions by drawing offers from one of two bingo cages, each with a known distribution. One cage has offers uniformly distributed on integers 1 to 10, and another on integers a to $a+9$ with $a < 10$. Subjects are not told which cage is being used, but must learn which by searching. This design leads to the unusual property that a subject may stop with a lower offer than she would continue searching with, if the higher offer provides the information that offers are being drawn from a higher-valued distribution.

Results with this design were remarkably similar to those with known offer distributions. Although subjects did slightly less well relative to optimal than with known distributions, searches were consistently terminated slightly earlier than the optimal model predicts.

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 undersearch 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.

Each of these results indicates that agents do not search enough, and that their search pattern is consistent with undervaluing the information gained from further search. Unfortunately, they also leave us with little information about whether undervaluation might operate in the broader domain of experimentation problems, including those in which there is more than one uncertain alternative.

Explanations for Suboptimal Search

The search experiments gave rise to two sensible explanations for suboptimal search: risk aversion and unobservable search cost.

In Cox and Oaxaca's experiments, agents had to choose between getting an additional draw from the wage distribution and a fixed payment for the rest of the session. Because continuing to search for a better wage is riskier than stopping, risk averse agents have lower reservation wages than risk-neutral agents. Therefore, risk averse agents will appear to stop searching too soon. Cox and Oaxaca argue this is what causes suboptimal search.

Although risk aversion could contribute to undersearch, there is evidence against its being the only explanation. First, Schotter and Braunstein used a lottery procedure to induce 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.

Pratt, Wise and Zeckhauser propose that suboptimal search could be attributable to an unobservable search cost. Although it probably does not explain suboptimal search in laboratory studies (where there is no search cost), it could complement risk aversion in the field, where agents must incur a cost of calling or visiting different merchants to determine their price for the commodity they are seeking. Since this (unknown) cost was not accounted for in their analysis, it could explain the supraoptimal price variance Pratt, Wise and Zeckhauser observed.

However, Pratt, Wise and Zeckhauser's data also provide an indication that suboptimal search may be 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

undervaluation is not undervaluation *per se*, but rather an unintended consequence of a simple choice rule which is poorly calibrated to these problems.

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 who is not induced to continue searching by the possibility of better offers in the future may not be induced to experiment because most of the benefit from present experimentation accrues in the future. 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.

The similarity between search and experimentation problems arises because both require agents to solve a dynamic programming problem to determine their best next action. If some feature of the utility function, bias or difficulty solving the dynamic programming problem leads to undersearch, it is reasonable to expect it could be manifested in experimentation problems. This link is further strengthened by the robustness of search results to unknown offer distributions. One aim of this study is to test this intuition that the cause of undersearch also leads to welfare loss in experimentation problems.

Risk aversion and unobservable search cost are both appealing because intuition suggests they are factors in experimentation problems. Although behavior in search problems is broadly consistent with the predictions of these models, they were not carefully controlled for, and there are some minor phenomena in the data which are

inconsistent with their predictions. Since they can be easily controlled in the lab, risk aversion and experimentation cost are excellent alternative hypotheses for a simple experiment which tests whether or not suboptimal information valuation extends to the experimentation environment.

1.3 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 the extensions of experimentation problems discussed here. 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.

1.3.1 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.

Information Value

The key concept in bandit problems, and the one which will eventually be used to identify the causes of undervaluation, 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 experiments described in Chapters 3 and 4. They ask 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.

1.3.2 Bandit Notation

For simplicity, attention is restricted to two-armed bandits. Otherwise, the notation I use largely follows that of Berry and Fristedt (1985).

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, except where specified, 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. I will also consider the case where Q is binomial payoff distribution with an unknown mean θ, $\mathcal{D}$ is the set of possible binomial distributions, and $\mathcal{G}$ is a beta distribution with unknown parameters α and β.

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.

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 $A = (1, \delta, \delta^2, \ldots)$. 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 case where A is exponential, which will be denoted $(F, \lambda; \delta)$.

Berry and Fristedt (1985) characterize the set of discount sequences for which a bandit reduces to an optimal stopping 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* $t = 1, 2, \ldots$

$$\frac{\gamma_{t+2}}{\gamma_{t+1}} \leq \frac{\gamma_{t+1}}{\gamma_t} \tag{1.1}$$

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

Knowing this is important because optimal stopping problems are much better understood, and much easier to compute solutions for, than the general bandit problem.

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

$$W(F, \lambda; A; \sigma) = E_\sigma[\sum_{\tau=1}^{\infty} \alpha_\tau X_\tau] \tag{1.2}$$

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),

$$V(F, \lambda; A) = \sup_\sigma W(F, \lambda; A; \sigma). \tag{1.3}$$

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

$$V^F(F, \lambda; A) = \alpha_1 E[X|F] + E[V((X)F, \lambda; A^{(1)})] \tag{1.4}$$

$$V^\lambda(F, \lambda; A) = \alpha_1 \lambda + V(F, \lambda; A^{(1)}) \tag{1.5}$$

These expressions will be useful in understanding the value function.

Gittins indexes and Information Values

Two more quantities are useful for comparing the value of information among arms. The Gittins index, denoted $\Lambda(F, A)$, is the value of a known mean arm for which the agent is indifferent between selecting the unknown arm and the known mean arm in the current period. The information value, the present discounted expected value of additional payoffs attributable to the information gathered from experimentation, is the Gittins index minus the expected value of the arm.

1.4 A Simple Example

To develop an intuition for the competing forces of maximization of current payoffs and gathering information which may improve future payoffs, consider the following two-armed bandit. The mean of the first arm is \$0.00 with p=1/2 and \$1.00 with p=1/2, and there is no noise, so the distribution of payoffs is the mean with probability one. The second arm has a mean of \$0.50 with p=1, and it too has no noise, so the distribution of payoffs is the mean with probability one. Therefore, arm 1 pays \$1.00 each time it is chosen half the time, and \$0.00 each time it is chosen half the time; arm 2 always pays \$0.50 every time it is chosen.

Suppose that there are $T > 1$ periods, and discounting is negligible. Although both arms have the same expected value, arm 1 is strictly preferred to arm 2. This is because a choice of arm 1 will reveal the value of arm 1, and if arm 1 gives a payoff

of \$1.00, it can be chosen in the future, yielding a payoff of \$1.00 each time.

Formally, the present value of a choice of arm 1 is given by

$$\frac{1}{2}\sum_{t=1}^{T} 1.00 + \frac{1}{2}(0.00 + \sum_{t=2}^{T} 0.50) = \frac{3T-1}{4}. \tag{1.6}$$

This reflects the payoff from a strategy where arm 1 is chosen in the first period. If it pays \$1.00 in the first period, it will pay \$1.00 in every period, so it is optimal to play it in every period, yielding a total payoff of $\$T$. On the other hand, if it pays \$0.00 in the first period, it will pay \$0.00 in every period and it is optimal to switch to arm 2. Then the payoff of \$0.50 is realized in each of the remaining $T-1$ periods.

Compare this to the value of choosing arm 2.

$$0.50 + \frac{1}{2}\sum_{t=2}^{T} 1.00 + \frac{1}{2}(0 + \sum_{t=3}^{T} 0.50) = \frac{3T-2}{4} \tag{1.7}$$

This reflects the initial choice of arm 2, yielding \$0.50, and then proceeding with the (optimal) choice of arm 1, advanced one period toward the horizon.

The best an initial choice of arm 2 can do is 1/4 less than an initial choice of arm 1. Therefore, although they have the same expected value, arm 1 is a strictly preferred initial choice because it can provide information about what is optimal in future moves.

1.4.1 The Value of Information

This example can also be used to emphasize two additional concepts in bandit analysis. The first is the idea of a known-payoff equivalent to an uncertain arm: how much would arm 2 have to pay each time it was chosen for the the agent to be indifferent between it and arm 1?

$$\begin{aligned} T\lambda &= \frac{1}{2}\sum_{t=1}^{T} 1.00 + \frac{1}{2}(0.00 + \sum_{t=2}^{T} \lambda) \\ \lambda &= \frac{T}{T+1} \end{aligned}$$

This number is called a Gittins index, and it represents the value of a certain payoff at which the agent is indifferent between playing the bandit and getting a certain payoff in each period. Note that the agent should always demand more than the expected value of the bandit for a certain alternative, because if she selects the bandit in the next period and receives a low payoff, she can always switch to the certain arm; it is only if the uncertain arm pays more than the certain arm that she will continue selecting it.

The Gittins index and the information value, the difference between an arm's Gittins index and its expected value, are the focus of the analysis in Chapters 3 and 4.

The second idea this example illustrates is that of a stopping problem. A bandit has the stopping property when, if it ever becomes optimal to play a known mean arm, it is optimal to select the known mean arm in every remaining period. This is intuitive in the example above: the agent learns nothing by choosing arm 2 at time t, and so has no information in time $t+1$ which would warrant a different choice than in time t. Furthermore, the horizon is one period closer, so the agent would have preferred to choose arm 1 in the previous period because then he would have had an additional period in which to use the information.

The stopping property is particularly important because it makes bandits tractable. In any period, the stopping property allows easy assessment of the value of switching to a known mean arm: the discounted sum of the known payoffs. Since this is a fixed number, rather than a dynamic programming problem, valuation of switching to a known arm is much easier. Without it, in fact, valuing bandits and computing optimal strategies is exceptionally computationally burdensome. For this reason, the experiments presented here have been designed to have the stopping property. The stopping property underlies all of the bandit theory presented.

1.5 Models of Suboptimal Bandit Behavior

The results from search problems suggest that risk aversion and experimentation cost may contribute significantly to apparently suboptimal behavior in bandit problems. In addition to these models, which are eventually rejected as the primary cause of undervaluation of information, three other behavioral models of choice are considered.

The first model, hyperbolic discounting, is the focus of Chapter 3. Hyperbolic discounting attributes suboptimal information values to the discount sequence. The model considered deviates from a standard exponential discount sequence in two ways. First, the sequence of discount factors themselves reflects an extreme, relative to exponential, preference for maximizing the present payoff, at the expense of future payoffs. In the particular form considered, exponentially discounted payoffs after the current period are further discounted by a constant factor (less than one). This means that agents undervalue information because they heavily discount the future payoffs which benefit from present experimentation.

The second way in which hyperbolic discounting differs from exponential discounting is that agents are not time consistent. With a hyperbolic discount sequence, an agent will often choose a known mean arm in the present period, but make an implicit plan to become exponential, and thus to experiment, in the next period. However, time inconsistency implies that the agent will be hyperbolic again in the future period, and will not honor this implicit commitment.

The second model which may explain apparently suboptimal behavior is ambiguity aversion. In bandits, the mean of the payoff distribution is itself drawn from a distribution. Ambiguity refers to the variance of this second-order probability, the distribution of the mean of the payoff distributions. If agents do not reduce compound lotteries, increases in the variance of the second order distribution can affect agents' valuation of a choice independent of risk attitude.

Chapter 4 examines whether a preference for relatively certain alternatives rooted in ambiguity aversion leads agents to undervalue information. It tests the prediction that subjects will have lower Gittins indexes than an ambiguity neutral agent, but

also that they will pay more than an ambiguity neutral agent for information about the true mean of an ambiguous arm.

The third model formalizes the intuition that agents' limited computational ability restricts the complexity of the decision tree they consider. Horizon truncation assumes that, instead of computing the value of information by considering the entire decision tree to the end of the horizon, agents consider the decision tree a small number of periods into the future, and then simply add a factor to compensate for the periods they did not consider. This model is not considered fully because it does not generate sharp predictions, but it does generate sensible, testable restrictions on behavior in each experiment.

The hyperbolic discounting and ambiguity aversion models are developed in a way which warrants special attention. They are both based on assumptions about behavior derived from stylized facts from psychology. These stylized facts are not designed to apply only to particular problems, but rather to be building blocks of theories which describe behavior in sophisticated environments such as bandits.

The model of rational optimal behavior assumes agents to be exponential discounters, and to use Bayes rule to reduce compound lotteries in calculating expected payoffs. By replacing each of these assumptions with assumptions which are based on stylized facts from psychology, I hope to build a theoretically rigorous model of bandit behavior which accurately describes observed regularities.

The predictions of the models can be formally tested with data, and accepted or rejected based on the empirical performance of the predictions. Further, if the behavioral model performs better than the optimal model, because I know which assumptions of the optimal model I replaced, I have learned the psychological cause of agents' deviation from optimality. Knowing why agents do not make optimal decisions can inform the policy discussions and corporate strategy sessions targeted at helping agents improve their welfare.

1.6 Summary and Hypotheses

Despite being an economically significant decision framework, computational complexity and lack of experimental control has discouraged researchers from considering whether or not behavior in bandit problems is optimal. Evidence from a single study of bandits indicates that agents do appreciate that there is some value to experimentation, but it does not provide information about whether agents experiment optimally. Several studies of search problems, a special case of bandits, demonstrate that agents do not search enough; this suggests that agents may not experiment enough in bandits.

The literature presents two plausible explanations for suboptimal experimentation, risk aversion and experimentation cost. Additionally, hyperbolic discounting, ambiguity aversion and horizon truncation predict that agents will not experiment enough. The next three chapters each present a laboratory experiment designed to test the predictions of one or more of these theories. The final chapter interprets the results, and offers some thoughts about the implications of the evidence for and against each of these theories for agent welfare and public policy.

Chapter 2 A Bandit Experiment

The literature reviewed in Chapter 1 suggests that agents recognize that there is value to experimentation in bandit problems, but that they may still not experiment optimally because they are risk averse or there is some unobservable cost to experimentation. This chapter has two objectives. First, to determine whether the tendency to undervalue information observed in search problems extends to the broader domain of bandit problems. Second, in the event that agents do not experiment optimally, to test experimentation cost and aversion to the variance of the mean-conditional payoff distribution, one component of risk, as explanations for underexperimentation.

This chapter presents the results of some simple choice experiments. Although they were not originally designed to test for Gittins optimal play, and therefore do not provide the best test of optimality, they do provide significant insight into how agents decide among bandit arms. The next section provides a precise definition of risk, and distinguishes it from ambiguity. Section 2.2 presents some of the challenges posed by theory, and gives some insight as to why bandits have not been heavily researched to this point. Section 2.3 discusses the experimental design. Section 2.5 demonstrates some of the difficulties in analyzing bandit data using discrete choice models. Results based on a variety of other statistical techniques are presented in Section 2.6.

2.1 Risk

Although risk is an intuitive concept, formally testing risk aversion in an environment with such a sophisticated belief structure as bandits requires a precise definition.

In bandits, payoffs are drawn from a distribution Q, which is itself uncertain. Q is drawn from a set of possible payoff distributions $\mathcal{D}$, with a probability given by the measure $G(Q)$. Therefore, there are two sources of variance in the payoff: variance

of Q and variance of G. The variance of an agents' payoffs arises from the combined effect of these distributions, and is represented by the subjective payoff distribution, F.

Definition 2 *An arm F is **riskier than** an arm F' if the variance of F is greater than the variance of F'.*

F is determined by combining G and Q using Bayes rule. If G is atomic, so there is only one possible mean, F will coincide with Q with the mean prescribed by G. If Q is atomic, so the payoff from an arm is equal to its mean every time, F coincides with G until a payoff is received, when F is updated to reflect the perfect information contained in the payoff.

2.2 Multi-armed Bandit Theory

This chapter considers the case where agents must choose among several uncertain alternatives. From a theoretical perspective, this is challenging because the dynamic programming problem that determines the optimal strategy must consider what would happen if each alternative were chosen at each state. Furthermore, because the valuation of each alternative is itself recursive, this does not reduce to a stopping problem at any state. Therefore, the optimal strategy can be extremely difficult to compute.

For one common discount sequence, however, Gittins and Jones (1974) have shown that a k-armed bandit problem can be reduced to k two-armed bandit problems, each consisting of one arm from the k-armed bandit and a known value arm. The index of the i^{th} arm, $\Lambda(F_i, \delta)$, is a function which gives the value of the known arm at which the agent is indifferent between selecting known and unknown arms in the i^{th} two-armed bandit. The theorem says that the optimal strategy is to select the arm with the highest index at each stage.

Theorem 1 *(Gittins and Jones, 1974) Suppose the discount sequence is the infinite horizon geometric with $0 < \delta < 1$. Then the optimal selection in the $(F_1, \ldots, F_k; \delta)$*

bandit is given by an arm with

$$\Lambda(F_i, \delta) = \max_j \Lambda(F_j, \delta). \tag{2.1}$$

This is a considerable computational simplification because the two-armed bandit is a stopping problem: once it is optimal to choose the known value arm, it remains optimal to select it in every subsequent period. The stopping property is easy to understand in the infinite horizon geometric case because selecting the λ arm provides no new information, so the agent has the same information and same continuation discount sequence in the next period; since nothing is different, the same selection must be optimal. Therefore, the value of switching to the λ arm in each period has a simple, closed-form expression, $\lambda/(1-\delta)$.

Unfortunately, Gittins and Jones' result does not generalize to nongeometric discount sequences.

Theorem 2 *(Berry and Fristedt, 1985) Suppose A is regular with $\alpha_1 > 0$. If, for all $(F_1, \ldots, F_k)$, the optimal initial selection in the $(F_1, \ldots, F_k; A)$ bandit are those i for which*

$$\Lambda(F_i, \delta) = \max_j \Lambda(F_j, \delta) \tag{2.2}$$

then $A \propto (1, \alpha, \alpha^2, \ldots)$ for some $\alpha_1 \geq 0$, i.e., A is geometric.

Berry and Fristedt conjecture that this generalizes to discount sequences with $\alpha_1 = 0$ and nonregular discount sequences.

This result implies, among other things, that the n-horizon uniform does not have the index property. Therefore, the most experimentally convenient design is not available in multi-armed bandits. The designs in later chapters, which have two armed bandits with one arm known, can utilize the n-horizon uniform because one of the arms is known, and therefore the bandit is a stopping problem. However, testing the predictions of Gittins optimal theory in a multiple-alternative choice environment requires inducing a geometric discount sequence.

2.3 Laboratory Bandits

In the laboratory, it is possible to test for the effects of risk aversion and experimentation cost by controlling for them. Experimentation cost is easily controlled for in the laboratory: do not charge agents to switch arms in subsequent periods. This may not capture all forms of "experimentation cost," such as cognitive cost associated with changing or keeping track of the payoffs associated with different alternatives. However, it does capture the costs associated with switching that Pratt, Wise and Zeckhauser argue lead to undersearch.

One dimension of risk, variance in the mean-conditional payoff distribution Q, can be controlled for by making the variance of Q the same for all arms. For instance, $\mathcal{D}$ can be the set of normal distributions with unknown means μ and known variance σ^2. Therefore, subjects cannot reduce their mean-conditional payoff variance by switching from one arm to the next.

It is impossible to control for both aspects of payoff variance in the bandit environment because risk cannot be help constant from one period to the next. Whenever a payoff is received from an arm, the agent must update her prior over the possible means of the arm, decreasing the variance of F.

In this environment, where all alternatives are equally risky and there is no experimentation cost, subjects should behave optimally if mean-conditional risk aversion and experimentation cost are the major factors contributing to undersearch and underexperimentation.

2.3.1 Gittins Indexes in the Laboratory

As suggested in Section 2.2, the primary challenge when working with bandits is actually computing the indexes. This section discusses some of the tradeoffs which must be made to test for mean-conditional risk aversion.

The first restriction imposed on the experimental design by Gittins theory is that the discount sequence must be geometric. This can be induced with some random chance that the bandit will end after each period.

The second restriction is imposed by the desire to control for mean-conditional risk aversion. Perfect control is attainable if all arms are equally risky, then risk averse agents will not have a risk-based motive for preferring one arm over another. This rules out payoff distributions like the Bernoulli, because the variance depends on the same parameters as the mean. Control is easiest to achieve with a normal distribution because the mean and variance can be chosen independently. Therefore, the mean of each arm will be unknown, and drawn from a known distribution G. When a subject selects an arm, the payoff is determined by adding random noise from a known distribution with mean zero and variance σ^2.

Computing Gittins Indexes for Normal Arms

Although geometric discounting ensures that the two-armed bandit used to determine the index of each unknown arm is a stopping problem, the infinite horizon requires that the stopping problem be computed for a very large state space. In practice, infinite horizon dynamic programming problems are approximated by solving very long finite horizon problem using backward induction. The error in the value function associated with the approximation is decreasing in the length of the horizon and increasing in the discount rate. Therefore, given a fixed computational horizon, decreasing the discount rate will result in a better approximation. (See Berry and Fristedt, 1985, Section 2.6 and Sutton and Barto, 1999 for more detail.)

In some cases, there is a sufficient statistic which can reduce the problem to a tractable one; other times, the index must be computed for each sequence of histories. Gittins (1989, chapter 6) demonstrates that sufficient statistics do exist for the normal distribution, but only if G is an improper prior. If this is the case, the index of an arm which has been chosen n times with an observed mean of $\bar{x}$ and a known payoff standard deviation of σ is given by

$$\Lambda((\bar{x}, \sigma, n), \delta) = \bar{x} + \sigma\Lambda((0, 1, n), \delta). \tag{2.3}$$

The index is a simple function of the index of a standard normal arm which has been

chosen n times. Values of $\Lambda((0,1,n),\delta)$ for several n and δs are published in Gittins' book.

Unfortunately, it is probably not possible to induce a truly improper prior in the laboratory; subjects would not believe that a billion dollar arm were possible because we could not afford to pay it. However, it is possible to induce a proper prior in the laboratory. For a conjugate normal prior, Bayesian updating is relatively simple.

With an improper prior, the agent becomes a frequentist who updates the mean of the payoff distribution according to $m_{n+1} = (n * m_n + x_n)/(n+1)$. This updating rule is invariant to the specific value of m. However, if the prior is informative, the Bayesian updates

$$m_{n+1} = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x}_n + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu \tag{2.4}$$

where $\bar{x}_n$ is the average of observed payoffs through n trials (with $x_0 = 0$), σ^2 is the known variance of the payoff distribution and μ and τ^2 are the known mean and variance of the prior distribution.

Because this expression depends on μ, τ and σ, it is not possible to compute a single index value which can be used to compute the index at a variety of priors and observed payoff sequences. Instead, the index must be recomputed for each combination of prior and average after n observations. This is prohibitively computationally intensive.

There are five ways to proceed. First, another distribution, which does allow use of sufficient statistics could be used. Second, the induced discount rate could be very low, so there were few expected periods, and therefore a good approximation is obtained by computing the Gittins index using a relatively shorter finite horizon. However, then so few selections of the arms are observed that it is difficult to distinguish if agents eventually converge to the correct arm.

Third, the experiment could be designed such that the sequence of payments from each alternative were the same for each subject, then the mean after n selections of each arm is known; those are the only values for which it is necessary to compute Gittins indexes; this is still computationally intensive, but less so. However, this

reduces the ability to use statistical information on the optimal amount of being wrong to assess whether agents converge to the optimal arm at the optimal rate.

The two other alternatives both leverage the published tables of improper prior Gittins indexes in Gittins (1989). Third, agents could be provided with a prior and the Gittins index could be computed based on the number of periods of experience that best approximates the prior. However, the prior may not be well represented by any number of periods, and the speed of adjustment to the mean, and therefore the Gittins index, is different than that of a Bayesian. Fourth, subjects could be given little distributional information about the prior, and therefore an induced improper prior. That the arms could have any mean may not be credible, however, so the model would be wrong in the case where agents came up with their own (uncontrolled) prior. Worse yet, the model would predict higher information values and Gittins indexes for agents who had more information, complicating documentation that undersearch extends to the more general bandit environment.

2.4 Experimental Design

The experiment discussed here was not originally designed to test for Gittins optimal play. Therefore, there are some features of the design which could be improved for this purpose. Nonetheless, several interesting features of the data suggest regularities which do or do not correspond to optimal play.

There are two treatments in this experiment, one with an induced proper prior and one with no induced prior (an improper prior). In each treatment, subjects were asked to play a four-armed bandit. The payoff mean of each arm is unknown and drawn from some distribution G. Ex ante, the arms are identical; the subject must, through sampling, determine which arm has the highest realized mean and draw from it in each period in order to maximize her earnings. When a subject selects an arm, the payoff is determined by adding random noise from a distribution with mean zero and variance σ^2.

In the improper prior treatment, subjects were told very little about the distribu-

Value	5	6	7	8	9	10	11	12	13	14	15
Roll	1	2	3	4 5	6 7	8 9	10 11 12	13 14 15	16 17	18 19	20
Chance	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{1}{20}$

Table 2.1: Distribution G from which arm means were drawn

tion from which payoffs are drawn, in an attempt to induce an improper prior. They were told only that the variance of the payoff distribution is the same for each arm, and that it is symmetric and unimodal. Since they are told the payoff variance is the same on all arms, risk is controlled. They are told nothing about G. Subjects were slightly deceived in this treatment, as payoff distribution means were not in fact drawn from a uniform distribution over the real line. If subjects knew this distribution, they should have been able to use this information to perform better than (improper prior) optimal.

In the informative prior treatment, subjects were given a prior which approximated a discretized normal (with very large bins). The priors were represented in tables, one showing how G is determined, and one showing the distribution of noise that generates the payoff distribution. Because this distribution is the same for each arm, there is no risk advantage to switching arms. Data from this treatment will not be compared with optimality, since it is very difficult to determine what optimality is, given the experimental design.

In both treatments, Table 2.1 was the basis for the prior distribution. To prevent subjects in the improper prior treatment from learning about the true prior, simple transformations of this distribution were used in some bandits.

The infinite horizon is induced with a 4% chance that the series will end after each period. Each period was concluded with one subject drawing a marble, with replacement, from a bag of one clear marble and 24 blue marbles; when the clear marble was drawn, the bandit was ended.

Subjects played several bandits during each experiment. After the clear marble was drawn, new payoff means for each arm were determined, the change was an-

nounced to subjects, and the new bandit was played. Subjects played a total of 250 to 300 periods.

2.4.1 Subjects and Procedures

This experiment was run on Caltech undergraduates who, although they did not necessarily have any economics training, had participated in previous unrelated economic experiments. The improper prior treatment used nine subjects, and the proper prior treatment used eight subjects. Upon arrival, they were asked to sit at a private computer terminal. Once all subjects arrived, the monitors were turned on and they read a set of on-line instructions while the experimenter read them aloud. Subjects were then given a brief quiz over the instructions; responses to the quiz indicated they understood the instructions, their environment and how their payoffs were determined. The instructions for both treatment are given in Appendix 2.A.

After the quiz answers were reviewed, one subject publicly counted the 24 blue marbles and one clear marble and placed them in a bag. Then the first series, or set of periods leading to a draw of the clear marble, was begun. Each series started with the experimenter rolling a 20-sided die four times. For instance, in the first series, arm mean distribution Table 2.1 was used, and the die rolls were 2, 18, 7 and 2, so for the arms labeled Blue, Green, Red and Pink respectively, the means were 6, 14, 9 and 6, and the standard deviation for all four arms was 20. The experimenter entered the arm means and the arm distribution variance into a private monitor, instructed the subjects to reload the JavaScript program used to run the experiment and asked the subjects to make their first period's choice. After the first choice, the experimenter asked one subject to draw a marble from the bag. If the marble was blue, the marble was replaced and subjects were asked to make another choice and another subject drew another marble, and so on, until the clear marble was drawn.

Throughout the experiment, subjects had constant access, through panels on their screens, to the instructions and to their complete playing history for that run of periods. Upon the start of new run of periods, the experimenter rolled the die again

to select new arm means, chose a new arm variance and informed subjects that they were chosen with a different distribution and the chances of particular deviations had changed.

The analysis presented here will rely only upon the data collected in the fourth and fifth series in the improper prior treatment and the seventh and eighth series in the informative prior treatment. The others did not contain enough periods to effectively assess bandit behavior.

2.5 Analysis of Bandit Choice Data

The case of a normal payoff distribution with an improper prior is a particularly appealing one to study because Gittins (1989) has calculated values which allow simple calculation of Gittins indexes for any combination of observed payoff mean, known or unknown payoff variance and the number of observations. Specifically, Gittins showed that the index $\Lambda((\bar{x}, \sigma^2, n), \delta)$ is a simple function of $\Lambda((0, 1, n), \delta)$ given by

$$\Lambda((\bar{x}, \sigma^2, n), \delta) = \bar{x} + \sigma\Lambda((0, 1, n), \delta). \tag{2.5}$$

For the case where the variance of the payoff distribution is not known, σ is replaced by s, the variance of received payoffs, and different values of $\Lambda((0, 1, n), \delta)$ are calculated.

Gittins provides values of $\Lambda((0, 1, n), \delta)$ in his appendix, for a variety of values of δ and n, and for known and unknown payoff distribution variance.

This formulation of the Gittins index is also analytically convenient, because it gives a very simple interpretation of the concept of information value. One natural hypothesis is that agents behave myopically, and pick the arm with the current highest expected value. In this case, they would be using a choice index which is just $\bar{x}$. The value of information the subject was using could be captured with weight put on the $\Lambda((0, 1, n), \delta)$ term. Therefore, one could test whether an agent was using the optimal information value in making choices in the bandit with the model

$$Pr(Y_t = i) = \Phi(\gamma\bar{x}_i + \phi\Lambda((0, 1, n_i), \delta)) \tag{2.6}$$

where $\Phi()$ is the extreme value distribution specified by McFadden (1978). An optimal player would have estimated parameters $\gamma = 1$ and $\phi = \sigma$. A myopic player would have $\gamma = 1$ and $\phi = 0$. Therefore, ϕ can be used to measure the degree to which agents appreciate the value of information.

2.5.1 Simulation Study

Although this model is statistically and conceptually simple, two recent studies (Salmon, 1999; Blume et al., 1999) have cast doubt on the ability of structural models like the one in Equation 2.5 to recover the choice parameters actually used by subjects. Both studies were concerned about the ability of structural models of learning in games to distinguish among putative models which could have generated the choice data. The heart of their criticism was that, in many circumstances, all of the learning models predicted the same choice. When this is the case, identifying the structural parameters which distinguish the models proved statistically impossible.

These studies make two different points about the estimation problem. Salmon tests several common learning models, and tests Camerer and Ho's claim that experience-weighted attraction (1999) can disentangle belief-based behavior from reinforcement behavior. His results are largely negative, however, a close reading of his paper reveals a serious problem: the games he uses have mixed strategy equilibria, which makes it hard to compare learning models against even random behavior. This paper leaves open the question of how well adaptive learning models perform in environments with more power to discriminate among them. This demonstrates the first important feature of an experiment for econometric power: the environment must be such that models make distinct predictions.

Blume et al. use a different set of games to perform a similar Monte Carlo study of how well learning models can identify the parameters used to simulate the data. Their results are quite encouraging, suggesting that modest sample sizes, say 26 choices, from a modest number of subjects, say 6 or 10, is enough to get an accurate picture of the parameters used to generate the data. However, they also found that

increasing the number of subjects does not help identify a misspecified model. This is an important point, as it means that small numbers of subjects are sufficient to obtain nearly asymptotic results with these learning models.

These results may bear on bandit problems because, in many choices, the myopic and optimal models will make the same prediction (i.e., the alternative with the highest predicted probability of being chosen). This will occur when one arm has a substantially higher mean than the others, when one arm has yielded a high outlier payoff than the others, or after each arm has been selected several times and even the optimal information value is small relative to the mean payoff. In these circumstances, it may be very difficult to distinguish an estimated ϕ from its optimal value, or from zero.

2.5.2 Monte Carlo Study Design

One way to understand whether or not the discrete choice model is useful in understanding bandit behavior is to run Monte Carlo tests of the model in Equation 2.5 on simulated bandit data.

The Salmon and Blume et al. studies are primarily concerned with whether or not small numbers of subjects can be used to identify the learning rules used in a population. Salmon is primarily concerned with the learning rule itself, and varies the learning rule on datasets of 40 subjects over 40 periods, and conducts his study on four different games. Blume et al. seek to identify a lower bound on the number of subjects necessary to correctly identify the learning rule used by subjects. They use groups of 6, 24, 96 and 384 to see if they can correctly identify the parameters in a fixed learning rule (exploring two interesting special cases).

In this study, I am interested in whether or not McFadden's conditional logit model can be used to distinguish optimal strategies from myopic strategies in simple bandit problems with small numbers of subjects. I examine the effect of three features of the choice environment on the statistical power of McFadden's model. It looks at different numbers of arms, different discount factors and different subject pool sizes.

The simulated data is designed to reflect as closely as possible the data generated by the experiment described in Section 2.4. Subjects played a sequence of bandits with improper priors, but with the actual prior given in Table 2.1. The discount rate was induced with probabilistic continuation rule with no new bandits starting after the 100^{th} choice and a maximum of 150 choices (although simulated subjects considered the problem to be infinite horizon). All subjects in a simulated experiment played the same arms with the same true means and number of periods, but they observed different draws from the payoff distributions.

Each subject began each bandit with two choices of each arm in order to identify the mean and standard deviation. The optimal choice model does not allow for any error in these first two periods; the Gittins index is infinity, and making a third selection of any arm before making the second selection of all the arms is infinitely unlikely. Therefore, these first choices were not included in the data used in the logit analysis.

After each arm has been chosen twice, the subjects selected the arm with the highest Gittins index, plus an error. As specified by McFadden's choice model, the errors were distributed extreme value, with unit parameters.[1] This is a critical assumption of McFadden's model; making choices without errors, or with a different error distribution, leads to much different results.

2.5.3 Results of Simulation Study

Each treatment consisted of 250 simulated bandits. In each bandit, three hypotheses are tested. First, a likelihood ratio test compares the model with unrestricted γ and ϕ to the optimal special case, where ϕ is restricted to one. Second, a similar test compares unrestricted γ and ϕ to the myopic special case, where ϕ is restricted to zero. Finally, a likelihood ratio test compares the unrestricted model to the model which generated the data, γ and ϕ both equal to one.

[1] The mean of the error distribution is γ, the Euler-Mascheroni constant, and the variance is π^2. (Wolfram, 9/6/2000)

Number Arms	2		4		8	
	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic
Reject Opt	6.0	0.0	4.0	0.0	0.8	0.0
Not Reject Opt	93.6	0.0	90.4	0.0	57.2	0.0
Reject Truth	4.8		3.4		2.8	

Table 2.2: Percentage of simulated datasets on which optimal and myopic indexes were rejected, across different numbers of arms

Number of Arms

Table 2.2 shows the results of the three hypothesis tests for bandits with different numbers of arms, 12 subjects and a discount rate of 0.90. The first and second rows of the table form an outcome matrix for each treatment. The upper left cell shows the percentage of times both the myopic and the optimal (true) models are rejected; the upper right cell, the percentage of times the optimal model is rejected but the myopic model is not; the lower left cell, the percentage of times the myopic model is rejected but the optimal model is not; and the lower right, the percentage of times neither model is rejected. If McFadden's choice model is statistically powerful in a given treatment, most of the outcome mass should be in the lower left cell.

The numbers in these cells need not sum to one hundred because there is a third outcome which is difficult to assess statistically, but which is important to consider when designing environments to test models of bandit choice. The third outcome corresponds to a computational problem which precluded performing some or all of the statistical tests of interest. In these cases, more than 1000 iterations of the Newton-Raphson search algorithm would have been required to locate the maximum of the likelihood function. With these data, the search algorithm either converges in less than ten iterations, or is very likely never to converge at all. Nonconvergence seems to be caused by data which does not provide sufficient information to identify the model, so the likelihood surface is very flat.

The third row of the table presents the percentage of simulated bandits in which the true model, with both parameters restricted to one, was rejected in favor of the

Discount Rate	0.80		0.90		0.96	
	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic
Reject Opt	4.0	0.0	4.0	0.0	1.2	0.0
Not Reject Opt	90.4	0.0	90.4	0.0	33.6	0.0
Reject Truth	3.8		3.4		4.6	

Table 2.3: Percentage of simulated datasets on which optimal and myopic indexes were rejected, across different discount rates

model with two free parameters.

In this study design, McFadden's model performs well with small numbers of arms. However, with eight arms, it actually arrives at parameter estimates less than 60% of the time. This is because the initial sampling takes 16 periods, and only after 16 periods is usable data produced. Bandits with significantly more than 16 periods are fairly unlikely with a discount rate of 0.90, so it is fairly common that there is little data on which to base estimates.[2]

Discount Rate

Table 2.3 shows how McFadden's model performs with different discount factors when there are four arms and 12 subjects. With lower discount factors, it performs very well as nearly all simulations converge, and nearly all that converge are able to detect the true parameters.

With the highest discount factor, however, relatively few of the simulated bandits converge. Unlike the convergence problems with large numbers of arms, the problem here is that the bandits have too many periods. After many selections of an arm, the change in mean and in information value after an additional selection is very small. Therefore, there is little switching, and little information which the model can use to identify the parameters. This suggests, that from a statistical standpoint, there is an optimal expected bandit length.

No. Subjects	8		12		25	
	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic	Reject Myopic	Not Reject Myopic
Reject Opt	4.4	0.0	4.0	0.0	4.0	0.0
Not Reject Opt	88.4	0.0	90.6	0.0	92.4	0.0
Reject Truth	3.0		3.4		4.6	

Table 2.4: Percentage of simulated datasets on which optimal and myopic indexes were rejected, across different numbers of subjects

Number of Subjects

Table 2.4 presents the results of McFadden's model for different numbers of subjects, with four arms and a discount factor of 0.90. This study suggests that the model is very powerful with a small number of subjects. The percentage of rejections of $\phi = 1$ is constant across the treatments. There is a very slight increase in the proportion of simulations which actually converge, but in each treatment, a considerable percentage do. Therefore, the conclusions of McFadden's model should be considered reliable, even with very small samples.

2.5.4 Implications of Simulation Study

Variation in each of the factors considered here leads to a risk that the data generated will not be useful in identifying the parameters of the logit choice model. This risk stems from two sources. First, there may not be a sufficient number of periods in each bandit, so very little usable data is generated. This occurs when the discount factor is so small, or the number of arms is so large, that there are rarely enough periods for the subjects to select each arm twice and then make subsequent selections.

Second, the Gittins indexes of the arms may be so disparate that it is difficult to identify just how much weight the subjects are placing on the information values. This occurs when there are few arms and the distribution from which the means are drawn has high variance. Then it is relatively likely that one arm will have a much larger mean than others, and the fewer arms there are, the more likely this

[2]The simulation did not count simulated experiments without any bandits of at least 17 periods.

difference is to be large. Similarly, if the discount rate is too low, then the bandits will be so long that little information is gained; after several periods of experience, the marginal information and marginal changes in information values have so little effect on Gittins indexes that there are few switches, and very little information to identify the information values subjects are using.

These trends are visible in the simulated data. When there are a large number of arms, bandits which last longer than 8×2 periods are sufficiently rare that there is not always enough data to identify the model. Similarly, when the discount rate is very high, there are not enough switches to identify the model. Therefore, in order to use the logit model, it is important to ensure that there is ample data and that there is a sufficient amount of switching within the sample to make it possible to identify the model; this study suggests that will not always be true.

The strong conclusion arising from this study is that it is very unlikely that a subject who is using an optimal strategy will be mistaken for a myopic subject. In fact, it never happened in any treatment in the study.

Also, conditioned on there being sufficient data for the model to be identified, the chance of being able to recover the true parameters is high. In no treatment did the frequency of rejecting the true parameters by a likelihood ratio test exceed the standard five percent threshold. Only when there were two arms did the frequency of rejecting true information value parameter restriction exceed five percent.

Based on this information, the primary concern for using McFadden's logit model on the data generated in this experiment should be that the discount rate is high, so there may not be enough switching in the available sample to identify the model parameters. In the actual experimental data, however, there are data from two bandits. At the beginning of each bandit, information values are high and estimated means can change considerably with one additional observation; therefore, there is a lot of switching, and a lot of information which can be used to identify the model.

Treatment	Series	Prior	Transform	STD	B	G	R	P	Periods
Improper	4	Table 2.1	+3	15	15	15	11	17	69
Improper	5	Table 2.1	None	20	8	6	5	14	85
Proper	7	Table 2.1	None	20	12	12	8	9	44
Proper	8	Table 2.1	None	20	8	11	12	13	69

Table 2.5: Parameters of each of the four bandits analyzed

2.6 Results

Since the experiments analyzed here were not specifically designed to test for optimal strategies, the focus in this section will be on identifying choice patterns which are and are not consistent with optimal and myopic strategies. Only the limited information treatment, where subjects did not know the prior distribution of the mean or the variance of the payoff distribution, will be compared to optimality, since there is no simple way to compute an optimal strategy in the informative prior case.

The distributions of arm means, realizations of arm means, arm standard deviations and number of periods in each bandit is presented in Table 2.5.

The data for each of the bandits are presented in Figures 2.1 and 2.2, along with the paths generated by simulating play of a large number of myopic agents. The data are grouped into four period blocks. Each series represents the proportion of choices in each block which were of the arm with each true mean. In each bandit, the arm with the highest true mean is a heavy black line; the arm with the second highest true mean is a light black line; the arm with the third highest true mean is a line with large dashes; and the arm with the lowest true mean is a line with small dashes. When two arms have the same true mean, their choice frequencies are averaged.

The choice paths of the simulated agents are represented by similarly-styled grey lines.

The first thing to notice is that arms with higher true means are played with greater frequency. Only in proper prior 8 is the highest mean arm not the most frequently chosen, and even then the arm which is most frequently chosen pays only one franc less per period.

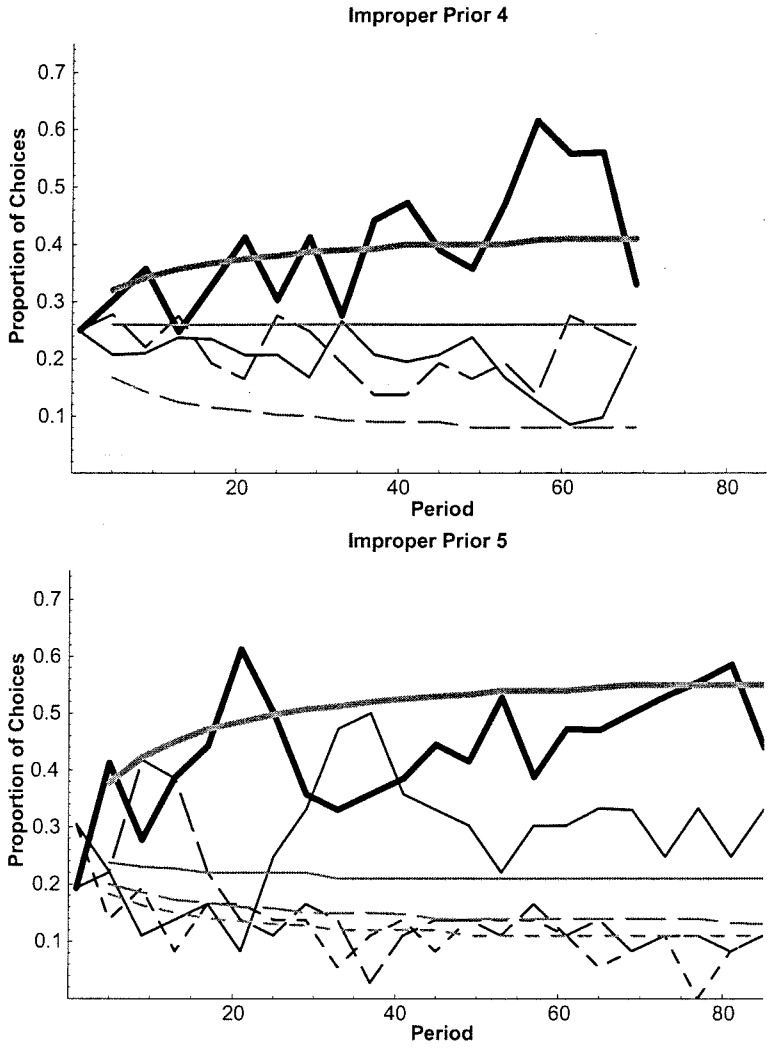

Figure 2.1: Arm choice frequencies, compared with paths generated by simulated myopic play, for improper prior treatment

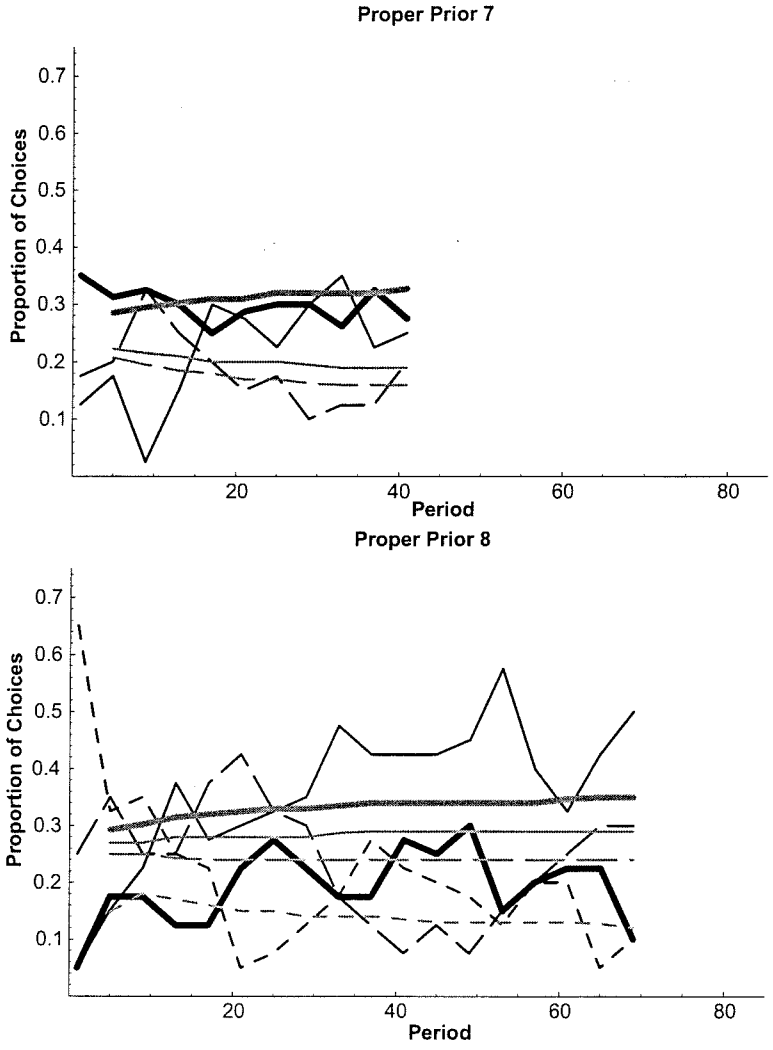

Figure 2.2: Arm choice frequencies, compared with paths generated by simulated myopic play, for proper prior treatment

This means that subjects are sensitive to payoffs, and do enough experimenting to detect which arms pay more. Furthermore, the frequency of playing the arm with the highest mean is increasing as time goes on, so subjects do not become insensitive to payoffs after several periods of experience.

The second thing to notice is that subjects choose the arm with the highest mean more frequently than the simulated myopic subjects. This suggests that subjects understand there is some value to experimentation, and thus do some experimenting to identify the best arm. Myopic subjects, on the other hand, lock in on the arm which yields the highest payoff the first time it is tried. The fact that they do not recognize the value of information acquired from experimentation means they choose arms with lower mean payoffs until their estimate of the mean drops below the observed mean of another arm.

Figure 2.3 shows the data from the improper prior treatment, along with paths generated from simulated optimal play. Although the subjects played the arm with the highest true mean more frequently than the myopic subjects, optimal subjects play the best arm yet more frequently. This suggests that subjects may not appreciate the full value of the information they gain by experimenting, and therefore may not experiment enough.

The remainder of this section identifies and statistically analyzes these patterns.

Result 1 *Arms with higher true means are played with greater frequency than arms with lower true means.*

Table 2.6 shows the proportion of periods (after the 30th) in which the arm with each true mean is the i^{th} most frequently chosen. When only three arms are reported, one is the average choice frequency of two arms with the same true mean. If two arms are chosen with equal frequency, each was assigned in the direction of its true rank (e.g., if the 17 and the 15 were both chosen 30% of the time, the 17 arm was ranked second and the 15 arm was ranked third).

The first thing to note in this table is that most of the mass is on the diagonal, so the arm with the highest true mean is chosen most frequently, the arm with the

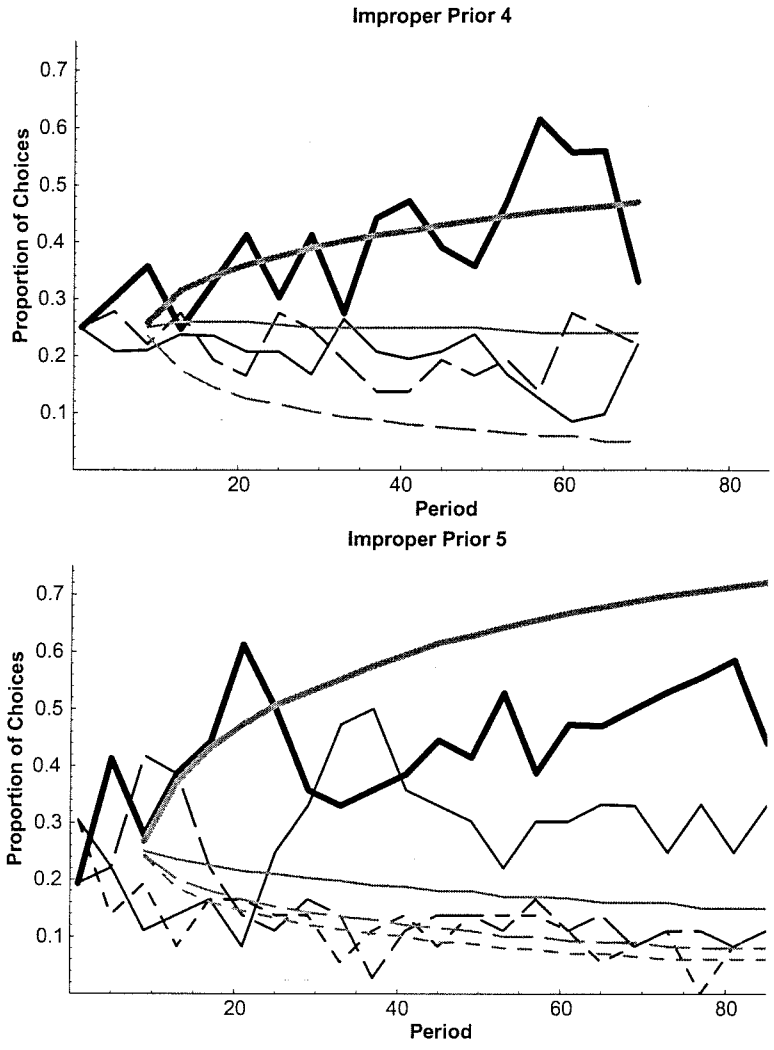

Figure 2.3: Arm choice frequencies, compared with paths generated by simulated optimal play

Improp 4				Improp 5				
Freq\Rank	17	15	11	Freq\Rank	14	8	6	5
1	0.85	0.10	0.05	1	0.74	0.21	0.05	0.00
2	0.13	0.51	0.36	2	0.23	0.56	0.12	0.09
3	0.02	0.39	0.59	3	0.03	0.13	0.64	0.21
				4	0.00	0.10	0.19	0.70

Prop 7				Prop 8				
Freq\Rank	12	9	8	Freq\Rank	13	12	11	8
1	0.55	0.25	0.20	1	0.14	0.66	0.14	0.06
2	0.43	0.48	0.10	2	0.34	0.22	0.31	0.14
3	0.03	0.28	0.70	3	0.40	0.06	0.29	0.25
				4	0.12	0.06	0.26	0.55

Table 2.6: Proportion of periods (after the 30th) in which each arm is the first, second, ... most frequently chosen arm

second highest true mean is chosen next most frequently, etc.

The second thing to notice about this table is that when there are deviations from the diagonal, they tend to be larger when the difference in means between two arms is smaller. In proper prior 8, for instance, the 12 arm is chosen more often than the 13 arm in several periods. Although this appears to be a significant error in terms of choice proportions, it is a small error in terms of payoff.

This implies that agents are sensitive to the true means of the arms and to the differences in the true means of the arms; subjects are eschewing random play for strategies which yield higher payoffs. This is consistent with both myopic Bayesianism and Gittins optimal play.

Result 2 *Average payoffs are consistent with both myopic and Gittins optimal strategies.*

Table 2.7 tests the hypotheses that, in each bandit, subjects' payoffs differed from the expected payoffs of playing either myopic or Gittins optimal strategies. Two values of subject payoff are considered: the actual average payoff received, and the expected value of the strategies the subject played. Using the expected value controls for the possibility a few large perturbations cause payoffs to differ significantly from

	Improper prior 4		Proper prior 7
	Optimal	Myopic	Myopic
Expected	1058	1057	467
Actual	952		318
p-value	0.10	0.11	0.00
Strategy Expected	1031		467
p-value	0.44	0.46	0.96
	Improper prior 5		Proper prior 8
	Optimal	Myopic	Myopic
Expected	919	874	792
Actual	876		818
p-value	0.55	0.98	0.44
Strategy Expected	844		766
p-value	0.23	0.62	0.17

Table 2.7: Actual and expected payoffs of observed strategies compared to expected payoffs of myopic and Gittins optimal strategies

the subjects' expectations.

The only significant difference in the table is in proper prior 7, where the actual payoffs are significantly less than would be expected if an agent were playing a myopic strategy. However, this is attributable to some bad realization from the payoff distribution, for the expected payoff from exactly the same sequence of arm choices is almost fifty percent higher, and just less than the myopic expect payoff. That such different results arise from actual and expected payoffs is further testimony to the difficulty of achieving statistical power in this environment.

In all four cases, the strategy expected payoffs are lower than those a myopic agent would receive. A Gittins optimal agent will do better than a myopic agent in expectation, so on average subjects are losing money relative to the optimal strategy. However, none of these differences is significant at conventional levels, even in the limited information treatment, when expected payoffs can be directly compared to those from an optimal strategy.

Therefore, looking at payoffs alone, it is not possible to determine whether subjects' behavior corresponds to myopic or optimal strategies. However, the structural model discussed in Section 2.5 may provide some insight.

Coefficient	Estimate	Std. Err.	z	95% CI	
Mean (γ)	0.217	0.010	21.46	0.197	0.237
Information Value (ϕ)	0.006	0.001	3.97	0.002	0.008

Table 2.8: Results of estimation of McFadden's choice model on improper prior 4 and improper prior 5

Result 3 *McFadden's choice model rejects both the optimal and myopic special cases.*

Table 2.8 gives the estimated parameters of the McFadden's choice model for the two improper prior bandits. Unfortunately, not all subjects began each bandit by selecting each arm twice, as the model requires to generate a prediction. Therefore, the data taken from each subject in each bandit begins after each arm has been selected two times. This may occur substantially later than the ninth period, and in one case, never occurred at all. These estimates are based on 958 choices.

The parameter of greatest interest is ϕ, for it indicates the extent to which subjects are considering the information values in their decisions. The estimates ϕ is significantly less than one, demonstrating that agents do not appreciate the value of information gained from present experimentation; they do not experiment enough. However, ϕ is also significantly higher than one, suggesting that subjects do appreciate that there is some value to the information gained from present experimentation.

The coefficient of the mean, γ, is also significantly different from both zero and one. However, it is much larger than ϕ, so the information value is not even in correct proportion to the mean.

Comparing the estimated model to the theoretical special cases with likelihood ratio tests also rejects both myopic and optimal strategies. The log-likelihood of the estimated model is -782.90. The optimal model has log-likelihood of -1328.07, leading to a test statistic of 1090.35 which is distributed $\chi^2(2)$, or a p-value very close to zero. The myopic model has a log-likelihood of -789.21, leading to a test statistic of 12.63 which is distributed $\chi^2(1)$, or a p-value of 0.0004.

Therefore, experimental subjects do recognize that there is value to information gained through experimentation, but consistently underestimate its magnitude.

Result 4 *When the myopic and optimal models make different predictions, subjects make the myopic choice more often than the optimal choice.*

Of the 958 observations in the improper prior treatment in which the models make predictions, the myopic and optimal models predict different choices in 669 cases. Of those cases, the optimal model is right 40 times, the myopic model is right 449 times, and one of the other two arms are chosen 180 times.

This pattern is consistent with subjects who do not appreciate the full value of information gained from experimentation. Most of the different predictions arise from cases where the information value on one arm is quite high, but its mean is lower than the mean of another arm. Subjects who do not appreciate the full information value will be drawn toward the choice predicted by the myopic model.

Once he has chosen the myopic choice, the information value of that arm continues to fall, while the index of the unchosen, optimal arm remains the same. The unchosen optimal arm will remain optimal with high probability, at least until it is chosen again. However, since the subject does not value information optimally, he may never choose it. Therefore, there are runs of many periods when the myopic and optimal models make different predictions, but the subjects repeatedly select only the arm predicted by the myopic model.

2.6.1 Sampling Patterns

Since subjects do not appear to be playing an optimal index strategy, the next natural question is: what are they doing? There are several regularities in the way subjects sample among their alternatives which express an understanding that information gained from experimentation has value, but which also are inconsistent with an index strategy. This section looks at the lengths of observed, optimal and myopic runs, as well as the number of distinct arms selected in the last five periods.

Runs

Figures 2.1 and 2.2 suggest that subjects do well initially, selecting each arm with equal probability, and then moving toward those which yield higher payoff. However, in each experiment, one high paying arm seems to be selected less often in favor of a low paying arm. One possibility is that one subject got a good observation on the lower-valued arm and is sticking with it as an index strategy would suggest. On the other hand, it could be that many subjects are continuing to select this arm after it should have been extinguished.

A run is when a subject selects the same arm several times in a row. Index strategies imply that as subjects get more information, they will converge on one arm, resulting in a long run. Figures 2.4 and 2.5 graph the average run length of experimental subjects, simulated myopic subjects and simulated Gittins optimal subjects against the period number for each bandit.

Result 5 *Subject run lengths are significantly shorter than those of a myopic agent.*

The simulated myopic agents, represented by a thick, dashed gray line, have fairly long runs. This is because switching requires the mean of one arm to drop below the mean of another. As more information is obtained, overcoming any given difference in means requires a progressively smaller (and lower probability) observation because of the diminishing impact of marginal information. They are not, as the Gittins optimal agents are, considering an information value which decreases as additional information is obtained.

Looking at Figures 2.4 and 2.5, the upper bound of the 95% confidence interval for the empirical average run length eventually falls below the average run length of the simulated myopic agents in every bandit. In the limited information bandits, average run lengths are never as long as those of a myopic agent; in the full information bandits, this happens relatively later, in period 20 in proper prior 7 and period 32 in proper prior 8.

This difference suggests subjects correctly perceive the need for greater information gathering in the limited information environment. The shorter run lengths mean

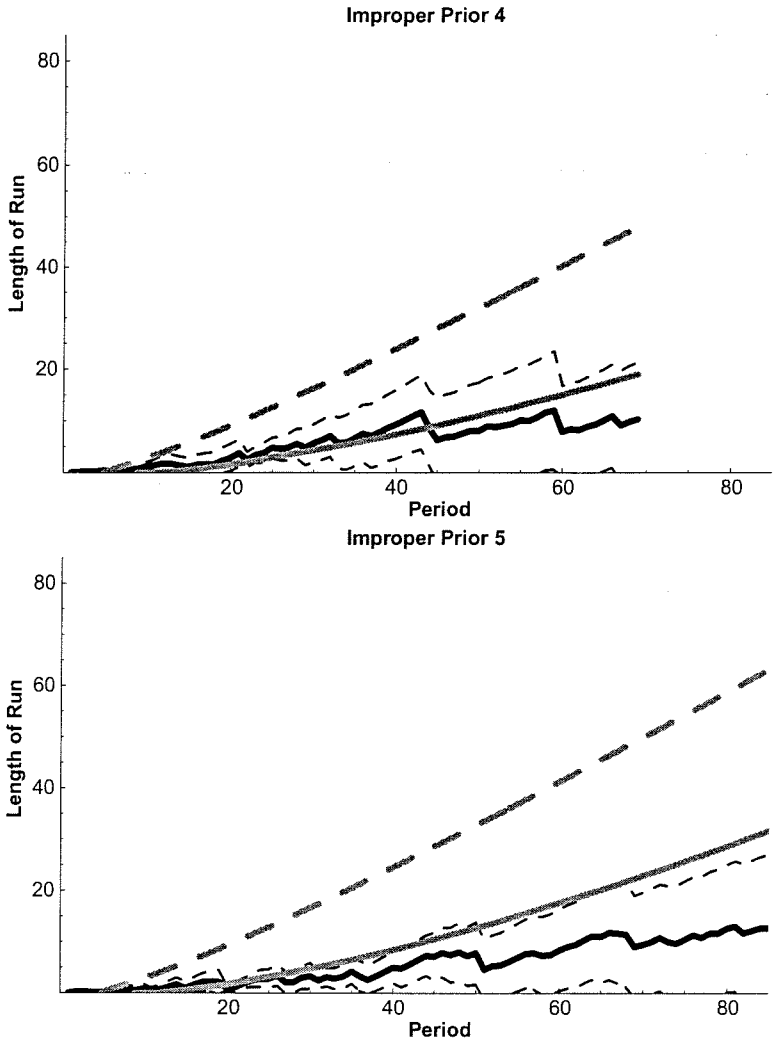

Figure 2.4: Average run lengths, compared with average run lengths of simulated optimal and myopic players, for improper prior treatment

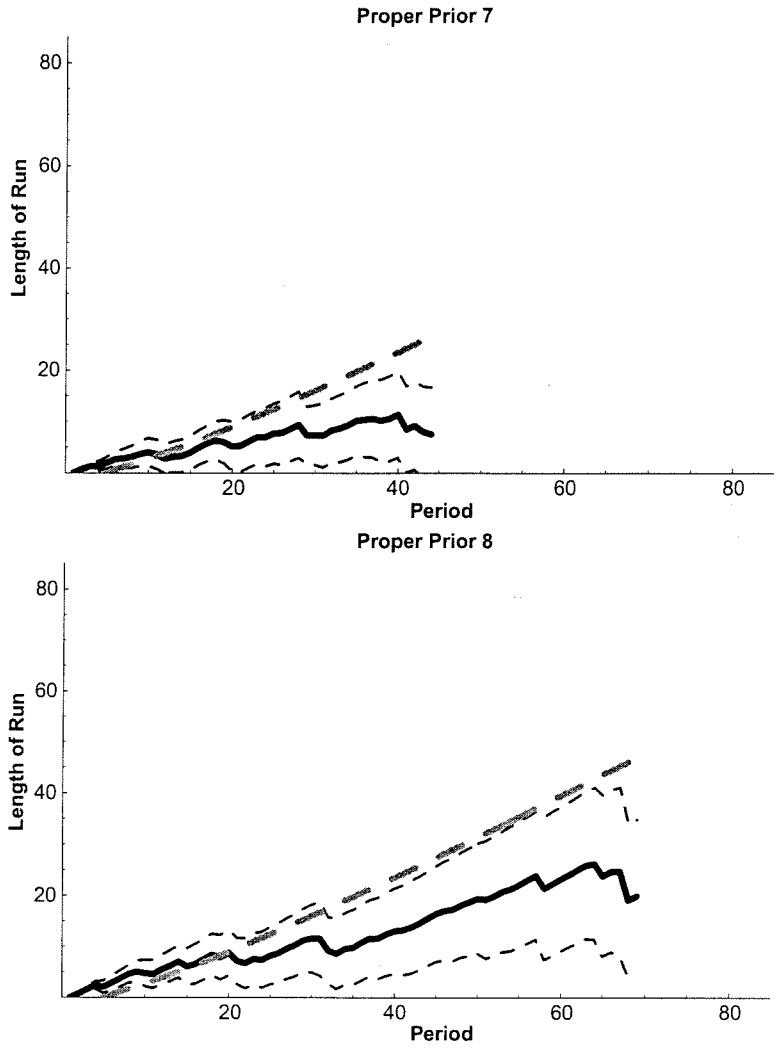

Figure 2.5: Average run lengths, compared with average run lengths of simulated optimal and myopic players, for proper prior treatment

they are switching among arms more frequently, gathering information rather than seeking to maximize current expected payoff. There is less switching in the full information bandits, as the prior provides information; however, subjects eventually do switch arms and experiment, resulting in shorter runs than myopic agents.

Result 6 *Run lengths are initially not significantly different from optimal run lengths, but are eventually shorter.*

The simulated Gittins optimal subjects are represented by a thick, grey line. In improper prior 4, the average run length tracks the optimal well until about the 45th period, when subjects begin switching more frequently than optimal. After period 60, the observed run lengths are only barely not significantly different than optimal.

The results in improper prior 5 are stronger, where the observed run lengths are significantly shorter than optimal in almost every period after the 30th.

That the observed run lengths are shorter than optimal means that subjects are switching arms more frequently than an optimal player would, especially in later periods. This may reflect that subjects are in fact experimenting, only that they are doing so after most optimal players will have converged on a single arm.

Intertemporal Mixing

Another dimension of switching behavior can be examined by looking at the number of distinct arms chosen in the last five periods. Figures 2.6 and 2.7 plot the moving average of the number of arms selected in the last five periods, along with 95% confidence intervals.

Result 7 *Subjects initially switch arms as frequently as simulated myopic agents, but continue switching arms after myopic agents stop switching.*

The path of simulated myopic agents is represented by a thick, dashed grey line. Myopic agents sample all the arms, but then quickly stop sampling because they play the arm which has yielded the highest average payoffs in each period. In each bandit,

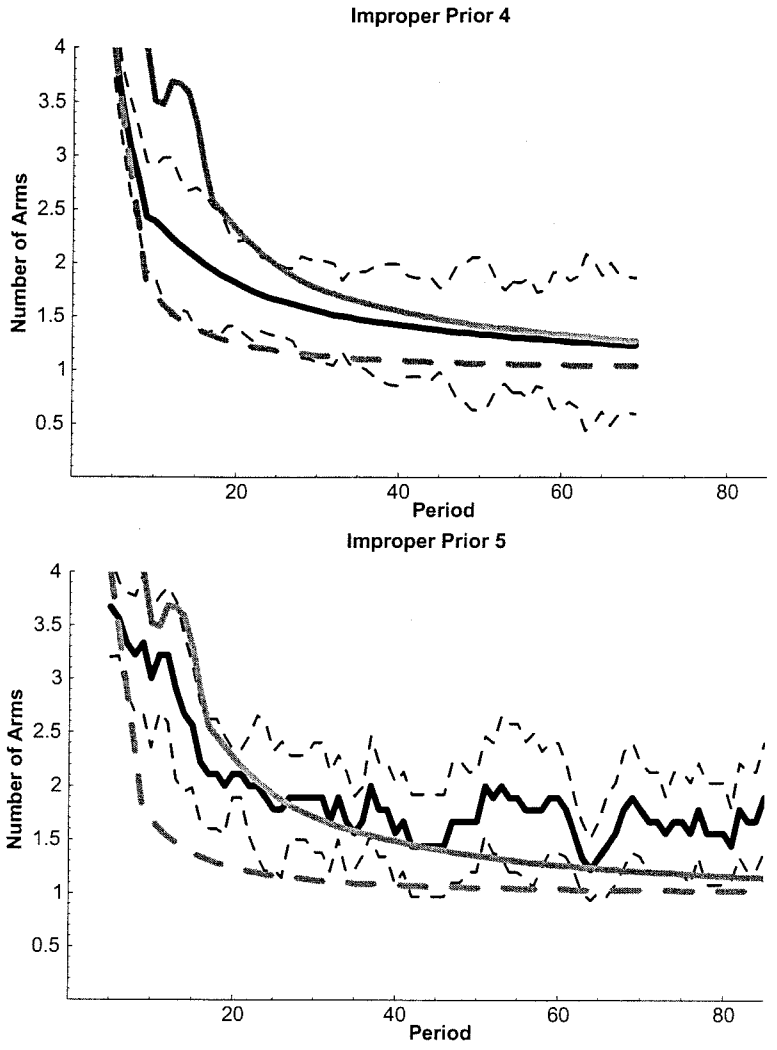

Figure 2.6: Average number of different arms chosen in the last five periods, compared with average number for simulated optimal and myopic players, for the improper prior treatment

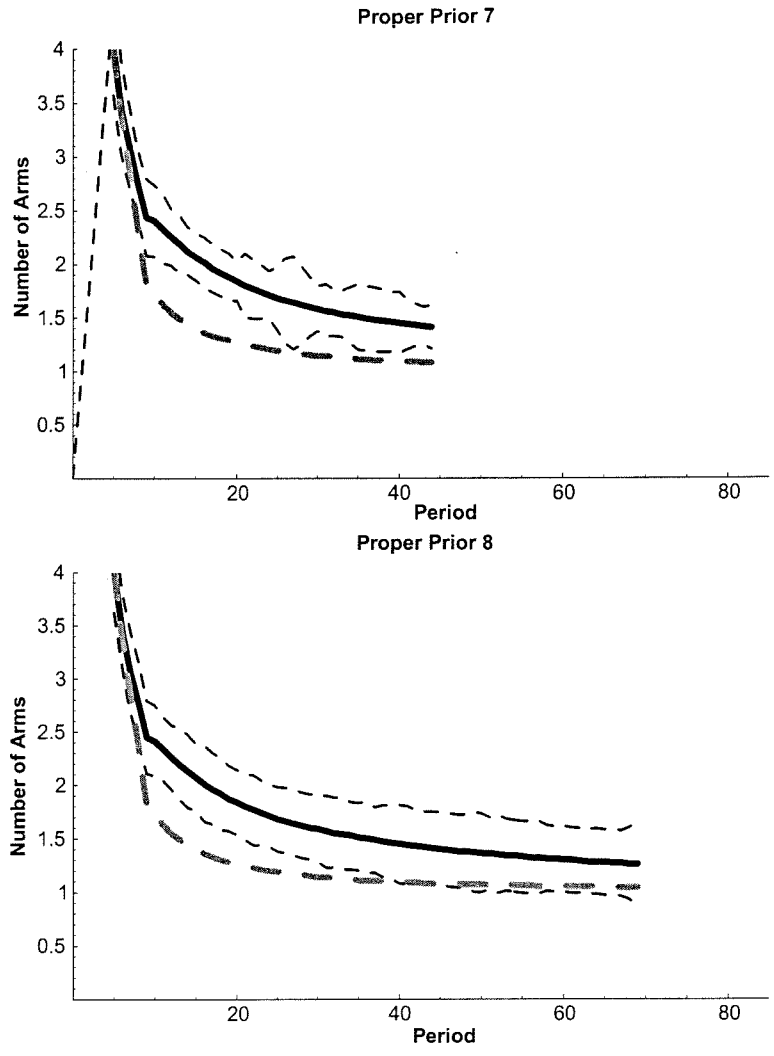

Figure 2.7: Average number of different arms chosen in the last five periods, compared with average number for simulated optimal and myopic players, for the proper prior treatment

on average, myopic agents had selected fewer than two arms in the last five periods as early as the eighth or ninth period.

In each bandit, the data initially follow the path of the myopic agents. The simulated myopic agents initialize themselves by selecting each arm once, so, after the fifth period, the myopic agents have selected all four arms in the last five periods. The subjects also appear to sample all the arms in the first several periods. However, between the fifth and tenth period, the rate of sampling drops off dramatically for the myopic agents. The data tracks this well for three or four periods as subjects avoid the arms which yielded poor first payoffs.

After this period of initial coincidence, the myopic agents continue to select few arms, but the experimental subjects do continue to experiment. This difference is statistically significant by the tenth period in every bandit. This statistical significant difference is maintained for fifteen or more periods; in improper prior 4 and proper prior 7, subjects are eventually selecting few enough arms that the myopic values again fall within the confidence bands; in improper prior 5 and proper prior 8, the myopic values never fall within confidence bands for a considerable number of periods.

Result 8 *Subjects do not initially sample as much as simulated optimal subjects, but they eventually sample at an indistinguishable or greater rate.*

The paths of simulated optimal agents are represented by a thick, grey line on the graphs for improper prior 4 and improper prior 5. Since these agents do not have priors, they must sample each arm twice to identify both the mean and variance of each arm.

In improper prior 4, the optimal agents sample a significantly greater number of arms than the experimental subjects. This difference remains significant for more than 35 periods, when the optimal agents converge enough to again drop within the confidence bands.

In improper prior 5, the optimal agents sample significantly more arms in periods five through sixteen. However, subjects' sampling levels off as optimal agents continue to reduce their level of sampling until the subjects' rate of sampling is not

distinguishable from optimality, and eventually is borderline significantly higher than optimal.

Considering the last four results together suggests a couple patterns in subjects' behavior. First, subjects do not appear to be using strategies entirely consistent with either optimal or myopic play. Rather, they appear to be engaging in some initial sampling, but then quickly dismissing those arms which do not pay satisfactorily. In improper prior 5, for instance, the 11 arm is dismissed after being chosen relatively few times. In these bandits, it happens that the arms they dismissed actually provided low payoffs, and thus their payoffs compared favorably with those of simulated optimal agents. However, this did not need to be the case. They often dismissed arms which gave initial bad outcomes when such outcomes could have been attributed to a bad draw from the payoff distribution, not a bad mean.

Once they have dismissed arms which have paid poorly initially, subjects experiment too much with the remaining arms; they continue to switch among arms, instead of trying to identify the one which yields the highest payoff. In improper prior 5, for instance, they switched freely and frequently among the 17 and two 15 arms. Because all paid well, they were able to do so with fairly little payoff penalty, but they were not maximizing their payoffs. In improper prior 4, where one arm is clearly more valuable than the others, there is much less such switching.

2.6.2 Initial Sampling

In the improper prior treatment, the Gittins model requires that each arm be selected twice, in order to identify the mean and standard deviation of each arm. In the informative prior treatment, however, such initial sampling is not necessary. However, many subjects select each arm several times, with relatively little sensitivity to payoffs.

Result 9 *subjects initially sample each arm a fixed number of times, with relatively little sensitivity to received payoffs.*

In the full information session, there is noticeable rule of thumb sampling from six of the ten subjects. The nature of the sampling is illustrated in Table 2.9. The

	Proper Prior 7			Proper Prior 8		
Sub	Samp Per	Samp Algorithm	No. Consist w. Sampling	Samp Per	Samp Algorithm	No. Consist w. Sampling
1	4	Cycle	4	4	Cycle	4
2	12	Cycle	12	69	Cycle	69
3	12	3 Cycle	12			
4	8	2 Cycle	8			
5	12	3 Cycle	10	7	Cycle	4
6	20	3 Cycle	12	12	4 Cycle	12

Table 2.9: Initial sampling patterns used by some subjects in the proper prior treatment

table shows the sampling outcomes for six subjects in two bandits, excluding the four subjects who did not exhibit sampling behavior in these two bandits (though they may have in other bandits). For each subject in each game, the table describes the length of the apparent sampling period where the rule of thumb describes behavior, as well as the number of those periods correctly predicted by the sampling algorithm.

The most common form of sampling is cycling, where subjects repeat the four arms in sequence several times; some subjects will choose each arm once before continuing play (ignoring their prior) and some will choose each arm four or five times before continuing play. For instance, eight periods of cycling might look like {1,2,3,4,1,2,3,4}.

Another common form of sampling is repeated arm sampling, where a subject will sample each arm a particular number of times before trying another arm. These are denoted in the table as N-cycles. A 2 cycle spanning eight periods might look like {1,1,2,2,3,3,4,4}.

As simple algorithms, these ways of sampling are not at all sensitive to the realized payoffs. In practice, subjects do respond to very good and very bad outcomes. Repeated arm samplers, in particular, will often sample an arm that has given a couple good outcomes one more time than the rule of thumb would predict, and an arm that has given a couple of very bad outcomes one fewer time than the rule of thumb would predict.

Unfortunately, these results do not have much predictive value; they are only useful for characterizing the nature of the suboptimality in subjects' strategies. Subjects

will often switch among the possible rules of thumb, and sometimes will be responsive to payoffs (especially when the payoffs are very good or very bad); I have been unable to recover any relationship between the rule used and payoffs received, even with respect to how long the sampling period extends. Thus, given that I have observed a subject using a rule of thumb on one bandit, I cannot predict her sampling behavior in another bandit.

That such sampling algorithms are used at all, however, indicates that subjects understand the need for sampling, though they do not use Gittins or index-based sampling strategies. Rather, there is a tendency for subjects to dispense with some sampling at the beginning of each period without significant sensitivity to payoffs.

2.6.3 Similarity-based Choice

Cycling is one strategy which does not rely on an index. Another is similarity-based choice, outlined by Rubinstein (1988, 2000). He claims that agents use judgments of similarity to eliminate dimensions of comparison on multidimensional choices. If two alternatives each have two dimensions and they are judged to be similar on exactly one dimension, then the alternative which is better on the dissimilar dimension is chosen. If one alternative is better on both dimensions, it is chosen; if the alternatives are similar on both dimensions, the model does not make a prediction.

This model could explain the pattern observed in the choice data in the following way. When information values are similar, that dimension is not considered, so the model would predict agents would select the arm with the higher expected value. This generates a different prediction than optimal theory if the arm with the lower expected value has a higher (but similar) information value.

However, after the arm with the higher expected value has been chosen several times, its information value falls and becomes dissimilar. The agent may then switch to the arm with the higher information value, leading to later overexperimentation.

This explanation is particularly appealing given that the indexes are probably not being explicitly computed; it makes sense for "close enough" judgments to be

important.

Result 10 *Rubinstein's similarity model predicts about as well as the optimal model on the subset (25-50%) of the sample on which it makes a prediction.*

Table 2.10 presents the difference in prediction rate of Gittins optimal theory and Rubinstein's model for pairwise arm comparisons for different notions of "small" mean differences and information value differences.

Since Rubinstein's model predicts only pairwise choices, the four-alternative choice is broken down into three pairwise choices, where it is assumed the chosen alternative was chosen over each unchosen alternative in a pairwise comparison; certainly it should never be the case that one alternative is preferred to the chosen alternative in pairwise comparisons.

There are a total of 2874 pairwise comparisons, of which 371 are excluded from analysis because one arm dominates the other (larger mean and larger information value). Of the remaining choices, the optimal model makes a prediction on all of them; the number of choices which similarity predicts varies with the value of "small". However, for the values considered here, most cells in the table represent predictions on 20-30% of the sample, with a maximum of 56% in the cell where the mean difference is 0.2 and the information value difference is 10. The optimal model correctly predicts 57% of total choices, more than the similarity model predicts at all.

Notice that the difference for every cell on the diagonal is zero. The reason for this is that, on the diagonal, both optimality and similarity make the same predictions. For a given value of the mean difference for which means are similar, s_m, and a given value of the information value difference for which information values are similar, s_i, the similarity model makes a prediction in two cases. First, it selects observations when the mean difference is less than s_m and the information value difference is greater than s_i, and predicts the arm with this higher information value will be chosen. The optimal model makes the same prediction when $s_i = s_m$, since the mean difference must be less than s_m, but the information value difference is greater than s_i.

Second, the similarity model selects observations when the mean difference is

Mean Difference	Information Value Difference						
	0.2	0.4	0.6	1	1.5	2	2.5
0.2	0.000	-0.014	-0.011	-0.005	-0.012	-0.028	-0.029
0.4	0.000	0.000	-0.008	-0.009	-0.014	-0.026	-0.029
0.6	0.051	0.003	0.000	-0.003	-0.008	-0.010	-0.007
1	0.121	0.045	0.000	0.000	0.000	0.001	0.004
1.5	0.223	0.112	0.007	0.000	0.000	0.000	0.003
2	0.264	0.146	0.026	0.006	-0.002	0.000	0.001
1.5	0.280	0.171	0.061	0.024	0.010	0.002	0.000
3	0.316	0.229	0.131	0.071	0.048	0.026	0.010
4	0.377	0.310	0.209	0.139	0.089	0.043	0.017
5	0.412	0.361	0.274	0.209	0.144	0.074	0.024
6	0.412	0.365	0.284	0.226	0.150	0.077	0.025
8	0.434	0.396	0.323	0.265	0.164	0.085	0.032
10	0.431	0.397	0.331	0.272	0.171	0.099	0.048

Mean Difference	Information Value Difference					
	3	4	5	6	8	10
0.2	-0.030	-0.028	-0.025	-0.031	-0.032	-0.038
0.4	-0.030	-0.027	-0.025	-0.032	-0.032	-0.039
0.6	-0.009	-0.008	-0.008	-0.016	-0.018	-0.026
1	0.000	0.000	-0.001	-0.010	-0.012	-0.021
1.5	-0.001	-0.001	-0.001	-0.010	-0.013	-0.021
2	0.000	0.000	0.000	-0.009	-0.014	-0.023
1.5	-0.001	0.000	0.000	-0.007	-0.009	-0.018
3	0.000	0.000	0.000	-0.009	-0.010	-0.017
4	0.000	0.000	0.000	-0.003	-0.003	-0.008
5	0.004	0.000	0.000	0.000	0.000	-0.001
6	0.004	0.000	0.000	0.000	0.000	0.000
8	0.010	0.006	0.003	0.000	0.000	0.000
10	0.024	0.014	0.012	0.006	0.002	0.000

Table 2.10: Difference between percentage correctly predicted by the optimal model and by Rubinstein's similarity model, for different values of "similar"

greater than s_m and the information value difference is less than s_i, and it predicts the arm with the greater mean will be chosen. The optimal model makes the same prediction when $s_m = s_i$ because the mean difference is greater than s_m and the information value difference must be less than s_i. Therefore, Rubinstein's model will not predict differently than the optimal model when $s_m = s_i$.

However, when $s_m \neq s_i$, an interesting pattern emerges from the data which, consistent with previous observations, suggests agents are undervaluing information. Consider the region where $s_m > s_i$. Two types of pairwise choices are selected into cells in this region. First, those where the difference in means is larger than s_m and the difference in information values is smaller than s_i. The similarity model predicts the arm with the larger mean will be chosen. The optimal model makes the same prediction, since the difference in means swamps the difference in information values.

The other observations in cells in this region are those where the difference in means is smaller than s_m and the information value difference is greater than s_i. For observations where the information value difference is larger than s_m, both the optimal model and the similarity model predict the arm with the larger information value will be chosen. However, for observations where the information value difference is between s_i and s_m, the two models make different predictions: the similarity model predicts the arm with the larger information value, and the optimal model predicts the arm with the larger mean.

The cells the region where $s_m > s_i$ in Table 2.10 are mostly positive, often considerably so. This means that in this region, the optimal model is predicting observed choices better than the similarity model; the arm with the larger mean is being chosen more often than the arm with the larger information value when the two are different. This is consistent with subjects undervaluing information.

This argument follows symmetrically for the region where $s_i > s_m$. Here, the optimal model predicts slightly worse than the similarity model. This is because different predictions are generated only on observations where the information value difference is less than s_i and the mean difference is greater than s_m. When the mean difference is less than s_i, similarity predicts the arm with the greater mean will be

chosen, but the optimal model predicts the arm with the greater information value will be chosen. Similarity performing better implies that the arm with the greater mean is chosen more often, which is consistent with subjects undervaluing information.

Analyzing data with respect to similarity confirms the earlier observation that subjects undervalue information. However, because it makes a prediction on only a small fraction of the data, it proves not to be a useful tool for analyzing behavior in bandit problems.

2.7 Discussion

Although the sample discussed here is small, and the data was originally collected for another purpose and not ideally suited for testing the predictions of Gittins theory, it provides some valuable insight into patterns of behavior which may appear in naturally occurring bandits.

These experiments indicate that undersearch does generalize to the more general bandit environment. Time series indicate that top-valued arms are discovered at a less than optimal rate because subjects do not experiment enough. This can be seen in the number of distinct arms chosen in the past five periods, where, early in the series, subjects choose among fewer arms than optimal samplers.

It is also apparent in analyzing the choice data with McFadden's choice model. Section 2.5 showed that this model is conclusive, even in fairly small samples. Estimating its parameters on the improper prior treatment data shows that subjects put less weight on the information value than do optimal samplers.

Significantly, this underexperimentation occurs even in this carefully controlled choice environment, where there is no explicit cost to experimentation, and where all alternatives are equally risky (conditional on the mean). That underexperimentation is still observed demonstrates that neither of these factors suggested in the search is the primary cause of suboptimality in bandits. Given this, one may also question their role in search problems as well.

Although subjects do not experiment optimally, they do not experiment myopi-

cally, either. The time series suggest much higher rates of play for higher mean arms than would be expected of myopic subjects. Further, the rate of switching to higher mean arms is greater than among myopic subjects. Even McFadden's model rejects the myopic special case.

These results suggest that subjects do understand that there is value to the information obtained from experimentation, but also that they do not appreciate its exact form. As demonstrated by the initial sampling patterns, subjects do initially sample from all the arms, but they curtail this sampling relatively sooner than a Gittins player. This means they focus on a single arm sooner than optimal, so they lose money in expectation because they are less than optimally confident that that arm provides the best combination of payoffs and information.

As they gain experience, however, subjects do not curtail their experimentation. Gittins optimal players switch arms fairly infrequently after many periods of experience, but subjects keep up a fairly constant rate of experimentation. Therefore, subjects are able to converge to the optimal rate of play of each arm.

The overall impression given by this data is that subjects are aware of the value of experimentation, but perhaps do not recognize that the value arises from the ability to use it in the future. Therefore, they do experiment, but not enough initially and too much after a number of periods. They could do better by doing most of their experimenting early; then they would have more time to exploit the information they gather.

Ideally, these results could easily be verified in an environment better designed to test for optimal strategies. However, there are several significant obstacles to doing so. Controlling even just for mean-conditional risk requires that all arms have a fixed, known standard deviation. This is most easily done with normal payoff distributions. However, computing Gittins indexes for a normal payoff distribution with an informative prior is very time consuming, and improper priors are not credible. Controlling for variance of the subjective payoff distribution is more difficult still.

Although these results need to be more carefully documented, in an environment with a better notion of optimality, many of the patterns observed here should be

robust to a cleaner experimental design. Surely agents will still engage in initial sampling, and this sampling will reflect an understanding of the value of experimentation.

The next natural question is, if underexperimentation is not attributable to risk aversion and unobserved experimentation cost, what causes the suboptimalities we observe? An ideal model would provide some insight into why subjects do not experiment optimally, and allow us to predict when underexperimentation was likely to be a problem. With this understanding, policies targeted at helping agents gather information could be developed. The next two chapters develop behavioral models which may provide insight into the psychological features leading to underexperimentation.

2.A Instructions

2.A.1 Improper Prior Treatment

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

The type of currency used in this experiment is francs. All transactions will be in terms of francs. Each franc is worth 0.008 dollars to you. At the end of the experiment, your francs will be converted to dollars at this rate, and you will be paid in dollars.

All communication during the experiment will be done through your computer terminal. The experiment will proceed as a series of periods during which you will make decisions and obtain earnings.

Urns

In each period you will be selecting an urn from which to receive a payoff. The payoff given by each urn in each period consists of two elements, a fixed value and a random value. Each urn is assigned a fixed value at the beginning of the experiment, and this fixed value does not change during the experiment. In this experiment, there will be

four urns.

Fixed Values

The fixed value of each urn will be randomly determined by a roll of a 20-sided die which will be done by the experimenter. The value on the die will be converted, using a table, to a fixed value for an urn. For this treatment, you can know only that the experimenter will use the same die and the same table for each urn.

Random Values

Although the fixed values of the urns never change, you will receive different payoffs each time you select an urn because the random value, which is added to the fixed value to determine your payoff, changes each period.

The random value is determined each period in a particular way. In this treatment, you can only be told on the following about the random values:

1. The average of the random values is zero, and the chance of getting a random value which improves your payoff is the same as the chance of getting one which lowers it.

2. The chance of getting a random value that adds X to the fixed value of the urn is the same as the chance of getting a random value which subtracts X from the fixed value of the urn.

3. The chance of getting a random value with a large absolute value is never larger than the chance of getting a random value with a small absolute value.

4. The chance of getting any particular random value is the same in every period.

5. The chance of getting any particular random value is the same for each urn.

There is some (very small) chance that very large and very small payoffs will be realized. For this experiment, single-period gains and losses will be capped at 1000 francs.

Using the Computer to Choose an Urn

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.

The History Panel

The long vertical panel on the left will contain your playing history. For each period, it will show your choice and the payoff you received; recent periods will be added to the top of the list, though later periods will still be accessible by scrolling down.

The Information Panel

The top of the three panels on the right side provides you with information on the current period and your total payoff, in francs. It also shows your previous period's choice and the payoff you received (information also available at the top of the history panel). There is nothing in this panel for you to modify.

The Urn Choice Panel

The middle of the three right-hand panels is where you indicate your choice of urn in 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 generate a random value and calculate your payoff for the period.

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.

Stopping Rule

You will choose an urn and receive a payoff in each period. At the end of the period, the experimenter will determine if an additional period will be played. This will be done by asking a subject to select a marble from a bag of 24 blue marbles and one clear marble. If the marble is blue, there will be another period. Thus, the chance of

there being another period is 96%.

Summary

1. The experimenter will announce the beginning of the period.
2. You will make a choice of urn and indicate it on the computer.
3. The computer will generate a random value. The random values are chosen such that the chance of a particular random value is the same for every period and for every urn.
4. The random value will be added to the fixed value of 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 and raise your hand.
7. The experimenter will announce the end of the period.
8. The experimenter will ask a subject to draw a marble from a bag. There will be another period approximately 96% of the time.

Feel free to earn as much money as you can. Are there any questions?

2.A.2 Proper Prior Treatment

The following sections replaced the same sections above.

Fixed Values

The fixed values assigned to each urn will be determined in a particular way. For each urn, the experimenter will select a number between 0 and 99 from a random number table (you will have the opportunity to verify the values after the experiment). The

random number will determine the fixed value of the urn according to the Fixed Value Table you have been given:

If the Die Roll is...	The Fixed Value will be:
1	5
2	6
3	7
5 6	8
7 8	9
9 10	10
11 12	11
13 14 15	12
16 17	13
18 19	14
20	15

Random Values

Although the fixed values of the urns never change once they are determined, you will receive different payoffs each time you select an urn because the random value, which is added to the fixed value to determine your payoff, changes each period.

The random value is determined each period in a particular way. The Random Value Table you have been given shows the chance that the random value will be greater than a particular value. The difference in the chances between two rows is the chance that the random value will be between those values. For instance, the difference between 0 and 10 is 19%, so there is a 19% chance that the random value will be between 0 and 10.

This same information is contained in the graph on the back of the table you have been given. This is what a histogram of thousands of random values would look like, with the values on the x-axis and relative frequency on the y-axis.

Notice that random values selected in this way have the following important properties.

1. The average of the random values is zero, and the chance of getting a random value which improves your payoff is the same as the chance of getting one which lowers it.

2. The chance of getting a random value that adds X to the fixed value of the urn is the same as the chance of getting a random value which subtracts X from the fixed value of the urn.

3. The chance of getting a random value with a large absolute value is never larger than the chance of getting a random value with a small absolute value.

4. The chance of getting any particular random value is the same in every period.

5. The chance of getting any particular random value is the same for each urn.

Chapter 3 Hyperbolic Discounting in Bandits

The simple choice experiment in Chapter 2 establishes that agents do not experiment enough in bandit environments, consistent with intuition based on behavior in search problems. Further, by observing underexperimention in the absence of experimentation cost and differential mean-conditional risk, it demonstrated that these putative causes of undersearch do not explain underexperimentation.

This chapter explores whether the observed underexperimentation is attributable to hyperbolic discounting, a formal model of intertemporal choice. Hyperbolic discounters discount future payoffs more than exponential discounters. Since the expected payoff increase from experimentation occurs primarily in the future, hyperbolic discounters are less inclined to experiment because the future payoffs which benefit from present experimentation are not as heavily weighted in their intertemporal utility function; they opt instead to maximize their present period payoff by selecting the arm with the highest expected value.

In addition to a careful exposition of how hyperbolic discounting affects play in bandits, this chapter also introduces horizon truncation, a simple model which is intuitively appealing and serves as a challenging baseline for the more sophisticated behavioral model. A horizon truncater looks only a few periods down the tree of possible outcomes, rather than to the end of the horizon. Therefore, the potential benefits to present experimentation which accrue past the truncated horizon are not explicitly incorporated into the decision. The version of horizon truncation discussed here allows agents to add a "fudge factor" for the value of the periods omitted, but does not impose any restrictions on this factor.

Understanding how present bias contributes to behavior in these environments is critical to helping agents maximize their welfare. The remainder of this chapter is dedicated to establishing the role of present bias in experimentation problems.

3.1 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 have proven 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.

3.1.1 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

$$\mathcal{U}(x_t, \ldots, x_T) = x_t + \beta \sum_{\tau=t+1}^{T} \delta^{\tau-t} x_\tau \quad \forall\, t. \tag{3.1}$$

A discount sequence of this form is also known as $\beta - \delta$ preferences.[1] 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

[1] 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 (1999) for a discussion of the differences between hyperbolic and quasi-hyperbolic preferences.

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 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.[2] 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.

3.1.2 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

[2] 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.

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 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 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.

3.2 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 the role of hyperbolic discounting in bandits.

3.2.1 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.[3]

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.

Notation for Bandits with Hyperbolic Discounting

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, $(\alpha_2, \alpha_3, \ldots)$.

[3]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.

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.

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.[4]

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 Appendix 3.A.1.

[4]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.

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.

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 3 *(Berry and Fristedt, 1985) There exists an optimal strategy σ^* for all possible priors G on $\mathcal{D}$ and all possible discount sequences A.*[5]

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 $\times$ 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.

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.

Theorem 4 *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 $\lambda \geq \Lambda(F, A)$.*

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

Proof: Please see Appendix 3.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 Appendix 3.A.2.

For $\beta < 1$, Theorem 4 establishes existence. For $\beta = 1$, the discount sequence is regular, so the existence result for regular discount sequences applies.

3.2.2 Properties of the Dynamic Allocation Index

The main result of this section, which confirms the intuition that hyperbolic discounters will not value information much as exponential discounters is given by the next theorem.

Theorem 5 *$\Lambda(F, \delta) \geq \Lambda(F, (\beta, \delta))$ for $\beta \leq 1$, and with equality only if $\beta = 1$.*

Proof: The proof of this theorem leverages the fact that the geometric discount sequence is regular while the hyperbolic discount sequence is not, along with the equivalence between regular discount sequences and stopping problems.

From the definition of $\Lambda(F, (\beta, \delta))$, $\Lambda(F, (\beta, \delta))$ is optimal in the $(F, \Lambda(F, (\beta, \delta)); \beta, \delta)$ bandit. Using the fact the naive hyperbolic discounter believes she will be exponential

in the future, the value of this bandit is

$$V(F, \Lambda(F, (\beta, \delta)); \beta, \delta) = \Lambda(F, (\beta, \delta)) + \beta V(F, \Lambda(F, (\beta, \delta)); \delta^{(1)}). \tag{3.2}$$

Because the hyperbolic discount sequence is not regular, the agent must expect to switch back to the F arm at some point. Since the continuation is regular, the only point at which the agent could choose F is at the second stage, in the $(F, \Lambda(F, (\beta, \delta)); \delta^{(1)})$ bandit. Therefore, $\Lambda(F, (\beta, \delta))$ is not optimal in the $(F, \Lambda(F, (\beta, \delta)); \delta^{(1)})$ bandit.

The value of the $(F, \Lambda(F, \delta); \delta)$ bandit is

$$V(F, \Lambda(F, \delta); \delta) = \Lambda(F, \delta) + V(F, \Lambda(F, \delta); \delta^{(1)}) \tag{3.3}$$

since $\Lambda(F, \delta)$ is an optimal choice.

Since the exponential discount sequence is regular, $\Lambda(F, \delta)$ is again optimal for the continuation.

Since the continuation in Equation 3.2 is the same as that in Equation 3.3, except for $\Lambda(F, \cdot)$, it must be that $\Lambda(F, \delta) \geq \Lambda(F, (\beta, \delta))$. Note that if $\beta = 1$, then the discount sequence in Equation 3.2 becomes geometric and the two bandits are the same. Since $\Lambda(F, \cdot)$ is unique, the theorem holds with equality.

Since hyperbolic discounters value the future payoffs which benefit from present experimentation less than geometric discounters, they have lower than optimal Gittins indexes. Therefore, even if they are using an index strategy, they will not experiment enough. Measuring β can help determine whether or not Gittins indexes reflect hyperbolic discounting.

Measuring β

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 1.4 and 1.5 to solve

$$\beta = \frac{\ell - E[X|F]}{\delta(E[V((X)F, \ell; \delta)] - V(F, \ell; \delta))}. \tag{3.4}$$

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.[6] An alternative measure of β is the information value ratio. The information value ratio is

$$\mathcal{I}(F, \ell, \lambda^*) = \frac{\ell - E[X|F]}{\lambda^* - E[X|F]}. \tag{3.5}$$

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.

3.3 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.

[6] 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.

3.3.1 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 3.4 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 Appendix 3.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.

3.3.2 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.[7] Its mean is randomly selected from the same $N(1, 25)$

[7]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

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 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.

3.3.3 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.[8] 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.

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.

[8] 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.

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.[9]

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; "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.

[9] 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.

3.3.4 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.[10] In each case, the data retained from these subjects are not idiosyncratic.

3.4 Results

Figure 3.1 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 round are based on the

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

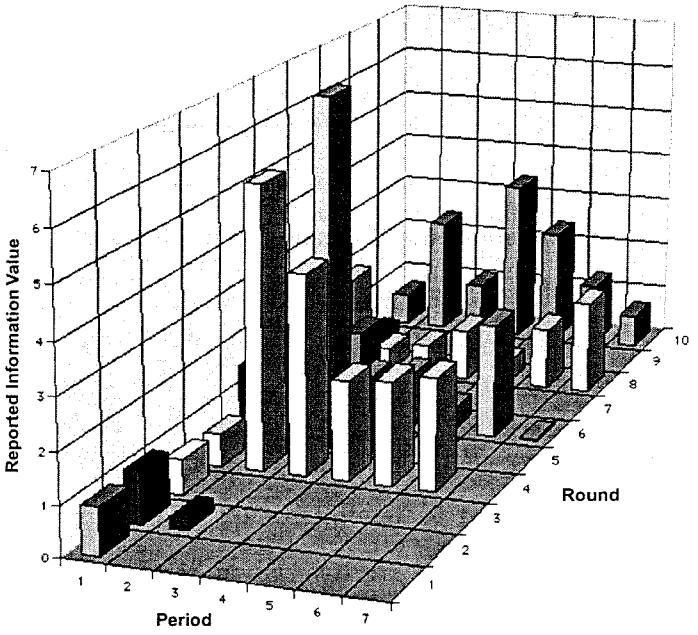

Figure 3.1: Information values reported by a typical subject

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 3.2 presents a box-and-whiskers plot of the information value ratio defined in Equation 3.5, 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 11 *First period ℓs are significantly below optimal, consistent with present bias.*

Support: Figure 3.3 shows the ℓs observed in the first period of each round.[11] This box-and-whiskers plot is interpreted the same as Figure 3.2, 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 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 3.1 presents the means, standard errors and the p-values for the one-

[11] 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.

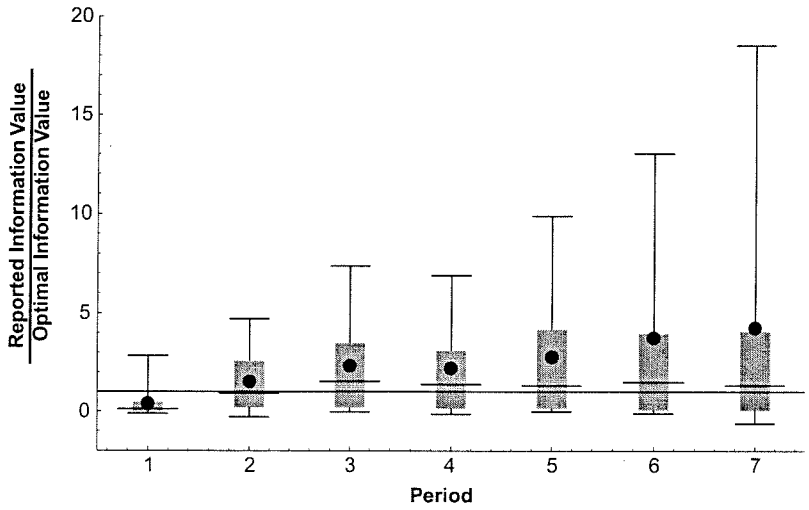

Figure 3.2: Box-and-whiskers plot of information value ratios across periods

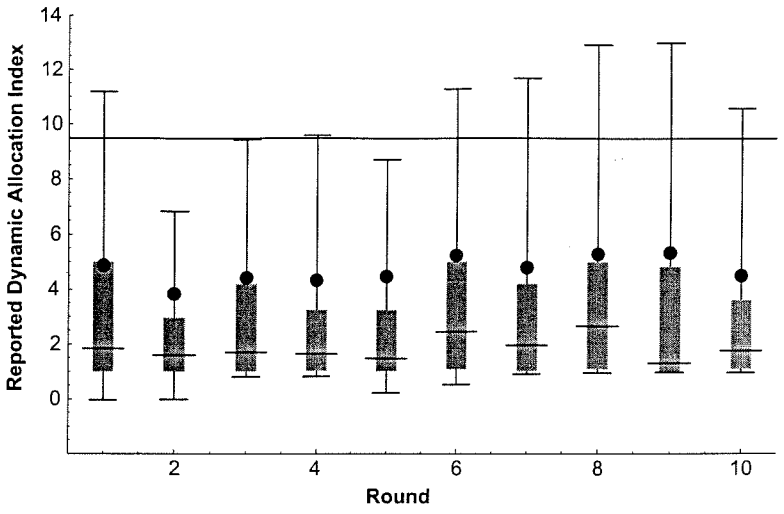

Figure 3.3: First period ℓs across rounds.

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 3.1: 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 3.2: Results of lag regression; the summary statistics are F(1,18)=109.11 and $R^2 = .728$

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 3.4 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 12 *The first period information values do not increase in later rounds.*

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

$$\mathcal{I}_t = \gamma_0 + \gamma_{lag}\mathcal{I}_{t-1} \quad \text{for} \quad t \geq 2. \tag{3.6}$$

Table 3.2 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}

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 3.3: Mean and median information value ratios for each period

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 13 *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 3.2, 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 3.3 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.

However, Figure 3.2 suggests the mean may not be the best description of the data. Although there is a clear upward trend in the mean, Table 3.3 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 estimate 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 3.4 presents the parameter estimates for the two segments, as well as the size of each segment. As Figure 3.2 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 is inconsistent with hyperbolic dis-

		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 3.4: Multicycle ECM estimates of two-segment regression model

Period	$\chi^2(1)$	p-value
2	35.55	0.000
3	20.41	0.000
4	20.05	0.000
5	17.51	0.000
6	6.94	0.008
7	1.78	0.182

Table 3.5: Test statistics comparing segment 1's median first period information value ratio to their medians in each other period

counting. This test is not strong enough to reject hyperbolic discounting for the rest of the population.

Table 3.5 presents a stronger test of whether or not the segment one subjects have increasing information value ratios. It compares the median information value ratio from the first period to the median information value ratios from other periods. This test counts the number of observations in each sample above and below the combined median and then computes a chi-squared statistic to determine the significance of the deviation (Seigel and Castellan, 1989).

The table shows that the median information value ratios in the second through sixth are significantly different (they are higher) than that of the first period; the insignificant result in the seventh period is most likely due to the small sample size. Therefore, even in segment one where information value ratios are not increasing after the first period, information value ratios are increasing across all periods in a way which cannot be explained by hyperbolic discounting.

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 3.6: Mean information values for each period for Rounds 5-10

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

Support: Table 3.6 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

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 3.7: 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 $R^2 = .083$

more responsive, so the difference is erased in later rounds.

Table 3.7 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.

3.5 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.

3.A Proofs of Propositions

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

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

$$\gamma_1 = 1 + \beta\delta + \beta\delta^2 + \ldots = 1 + \beta\delta \sum_{\tau=0}^{\infty} \delta^\tau = 1 + \frac{\beta\delta}{1-\delta}$$

$$\gamma_2 = \beta\delta + \beta\delta^2 + \ldots = \beta\delta \sum_{\tau=0}^{\infty} \delta^\tau = \frac{\beta\delta}{1-\delta}$$

$$\gamma_3 = \beta\delta^2 + \beta\delta^3 + \ldots = \beta\delta^2 \sum_{\tau=0}^{\infty} \delta^\tau = \frac{\beta\delta^2}{1-\delta}$$

Now plugging these into the definition of regular, we have

$$\begin{aligned}
\frac{\frac{\beta\delta^2}{1-\delta}}{\frac{\beta\delta}{1-\delta}} &\leq \frac{\frac{\beta\delta}{1-\delta}}{1+\frac{\beta\delta}{1-\delta}} \\
\delta &\leq \frac{\beta\delta}{1-\delta+\beta\delta} \\
1 &\leq \frac{\beta}{1-\delta+\beta\delta} \\
1-\delta &\leq \beta(1-\delta) \\
1 &\leq \beta
\end{aligned} \tag{3.7}$$

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. □

3.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 6 *(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;

$$\Delta(F, \lambda; A) = V^F(F, \lambda; A) - V^\lambda(F, \lambda; A). \tag{3.8}$$

The absolute value of 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

$$V^F(F, \lambda; A) = \alpha_1 E[X|F] + E[V((X)F, \lambda; A^{(1)})]. \tag{3.9}$$

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

$$V^\lambda(F, \lambda; A) = \alpha_1 \lambda + V(F, \lambda; A^{(1)}). \tag{3.10}$$

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

$$V^F(F,\lambda;A) = V(F,\lambda;A) - \Delta^-(F,\lambda;A) \tag{3.11}$$

$$V^\lambda(F,\lambda;A) = V(F,\lambda;A) - \Delta^+(F,\lambda;A). \tag{3.12}$$

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 3.9 and 3.10 above. This gives

$$V^F(F,\lambda;A) = \alpha_1 E[X|F] + E[V^\lambda((X)F,\lambda;A^{(1)}) + \Delta^+((X)F,\lambda;A^{(1)})] \tag{3.13}$$

$$V^\lambda(F,\lambda;A) = \alpha_1\lambda + V^F(F,\lambda;A^{(1)}) + \Delta^-(F,\lambda;A^{(1)}). \tag{3.14}$$

These expressions can then be used to compute $\Delta(F,\lambda;A)$. The first two terms in Equation 3.13 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 3.14 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 3.14 from those in Equation 3.13 gives $(\alpha_1 - \alpha_2)[E[X|F] - \lambda]$. This gives

$$\Delta(F,\lambda;A) = (\alpha_1 - \alpha_2)[E[X|F] - \lambda] + E[\Delta^+((X)F,\lambda;A^{(1)})] - \Delta^-(F,\lambda;A^{(1)}). \tag{3.15}$$

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

$$\begin{aligned}\Delta(F,\lambda^*;A) - \Delta(F,\lambda;A) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^+((X)F,\lambda^*;A^{(1)}) - \Delta^+((X)F,\lambda;A^{(1)})] + \\ \Delta^-(F,\lambda;A^{(1)}) - \Delta^-(F,\lambda^*;A^{(1)}).\end{aligned} \tag{3.16}$$

(ii) Proving the lemma requires that the difference in Equation 3.16 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

$$\begin{aligned} E[X|F] - \lambda^* &\leq E[X|F] - \lambda \\ \lambda^* &\geq \lambda \end{aligned} \tag{3.17}$$

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 3.16 can be rewritten with the truncated discount sequence

$$\begin{aligned} \Delta(F, \lambda^*; A_{n+1}) - \Delta(F, \lambda; A_{n+1}) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^+((X)F, \lambda^*; A^{(1)}_{n+1}) - \Delta^+((X)F, \lambda; A^{(1)}_{n+1})] + \\ \Delta^-(F, \lambda; A^{(1)}_{n+1}) - \Delta^-(F, \lambda^*; A^{(1)}_{n+1}). \end{aligned} \tag{3.18}$$

The first term on the right-hand side of Equation 3.18 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 $A^{(1)}_{n+1}$ is nonincreasing and has horizon n, we have

$$\begin{aligned} & E[\Delta^+((X)F, \lambda^*; A^{(1)}_{n+1}) - \Delta^+((X)F, \lambda; A^{(1)}_{n+1})] \\ = E[\max[0, \Delta((X)F, \lambda^*; A_n)] & - \max[0, \Delta((X)F, \lambda; A_n)]]. \end{aligned} \tag{3.19}$$

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 3.19 is always weakly larger than the first, implying that the second term in Equation 3.18 is nonpositive.

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

$$\begin{aligned} &\Delta^-(F,\lambda;A^{(1)}_{n+1}) - \Delta^-(F,\lambda^*;A^{(1)}_{n+1}) \\ &= \max[0,-\Delta(F,\lambda;A_n)] - \max[0,-\Delta(F,\lambda^*;A_n)]. \end{aligned} \tag{3.20}$$

The induction hypothesis gives that $\Delta(F,\lambda^*;A_n) \leq \Delta(F,\lambda;A_n)$. Therefore, the second term in Equation 3.20 is always weakly larger than the first, implying that the third term in Equation 3.18 is nonpositive.

Since each of the three terms in Equation 3.18 is nonpositive, we conclude that the difference Equation 3.16 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

$$\Delta(F,\lambda^*;A_\infty) \leq \Delta(F,\lambda;A_\infty). \tag{3.21}$$

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 3.16 is nonpositive. If $\alpha_1 > \alpha_2$, then the first term on the right-hand side of Equation 3.16 is strictly negative because $\lambda^* > \lambda$ by assumption. Therefore, Equation 3.16 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 4 *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

$$V^F(F, \lambda; A) > V^\lambda(F, \lambda; A) \tag{3.22}$$

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,

$$V^F(F, \lambda; A) < V^\lambda(F, \lambda; A) \tag{3.23}$$

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

$$V^F(F, \Lambda(F, A); A) = V^\lambda(F, \Lambda(F, A); A), \tag{3.24}$$

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 4 are met.

If $\beta = 1$, Berry and Fristedt's Theorem 5.5.3 applies directly, providing an exact analog of Theorem 4 for regular discount sequences. Since then the discount sequence A is regular in this case, the conditions of their theorem are satisfied.□

The existence of a dynamic allocation index should not be confused with the existence of an index result like that of Gittins and Jones (1974) 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.

3.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 3.3.1.*

Proof: This proof proceeds by showing that the mechanism of Section 3.3.1 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 3.3.1 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.□

3.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.

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.

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.

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 right-hand column shows the percentage of each of the one billion urns with each possible number on it. For example, 10,203,858 urns, or 1.023% of all the urns, have numbers between 17 and 18 on 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% of the urns have numbers between 11 (1+10) and -9 (1-10), and about 95% of the urns have numbers between 21 and -19.

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 right-hand column shows the percentage of each of the one billion balls with each

possible number on it. For example, 53,200,074 balls, or 5.32% of the balls in each urn, 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% of the balls have values between 5 (0+5) and -5 (0-5), and about 95% of the balls have values between 10 and -10.

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.

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.

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.

textbfThe 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.

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.

The Urn Choice Panel

Please look at the middle of the three right-hand 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.

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.

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?

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.

Chapter 4 Ambiguity Aversion in Bandits

Although agents' Gittins indexes are too low, the previous chapter demonstrated that they do not react to new information in the way predicted by hyperbolic discounting. Even the majority of subjects whose information value ratios converged to one as they acquired more information appeared to significantly increase their βs after their first observation.

A second behavioral model which may apply to bandits is ambiguity aversion. Ambiguity aversion holds that uncertainty about the mean of the payoff distribution leads to lower than optimal Gittins indexes and underexperimentation. Furthermore, as information about the mean of the payoff distribution is acquired, ambiguity is reduced and the observed Gittins index converges to optimality. Therefore, the apparently changing βs in Chapter 3 could in fact have been subjects reacting to the changing level of ambiguity as they learned more about the uncertain arm.

This chapter considers whether ambiguity aversion plays a role in bandits. It develops Kahn and Sarin's (1988) model of ambiguity aversion into a model of behavior in bandits with Bernoulli payoffs and beta priors. The formal model generates a seemingly paradoxical prediction: agents' Gittins indexes will be lower than optimal, so it appears they do not value information enough, but they will pay more than optimal for information about the value of an uncertain arm, so it appears they value information too much. This prediction is tested directly in a laboratory experiment.

4.1 Ambiguity Aversion

The concept of ambiguity aversion was most directly expressed by Ellsberg (1961). He posed a thought experiment where agents had to choose between an urn which contains 30 red balls and 60 balls in some unknown combination of black and yellow. Agents were asked to rank two pairs of bets: X gave a prize if a ball drawn from the

urn was red, and Y if the ball drawn was black; X' gave a prize if the ball is either red or yellow, and Y' gives a prize if the ball is either black or yellow.

Most people prefer X to Y and Y' to X'. This is paradoxical because this pair of preferences is inconsistent with any fixed belief about the mixture of black and yellow balls in their urn: choosing X over Y implies a belief that there are fewer than 30 black balls, and therefore more than 30 yellow balls which implies X' should be preferred to Y'.

The Ellsberg paradox demonstrates that people prefer known probability bets to unknown probability bets, and provides an operationalization of ambiguity. However, no consensus has emerged of a precise definition of ambiguity. Partly, this stems from the variety of circumstances in which different notions of ambiguity seem suitable. For instance, the credibility of a source of information, or the degree of disagreement between multiple sources, such as expert witnesses, captures one notion of ambiguity (Einhorn and Hogarth, 1985). Ambiguity may also represent uncertainty about probability stemming from information which could be known, but is not (Frisch and Baron, 1988). True subjectivists may reject the notion of ambiguity altogether, since all subjective probability distributions are equally well known to ourselves (deFinetti, 1977).

The notion of ambiguity considered here is based on second order probabilities. A second order probability is the distribution of possible distributions. In the Ellsberg problem, the second order probability is the probability distribution over the number of black balls in the urn. Second order probability is also commonly encountered as statistical confidence. Consider, for instance, two coins, one which has been flipped twice yielding one head and one tail and another which has been flipped 1000 times, yielding 500 heads and 500 tails. Both coins have $P(\text{heads})=0.5$, but the coin with more flips has a lower-variance second order probability.

This approach is not without its drawbacks. There is some evidence that agents prefer known second order probabilities to unknown second order probabilities (Yates and Zukowski, 1976), suggesting third order probabilities may also affect ambiguity. Also, if people have difficulty understanding second order probabilities, then higher

order probabilities are probably more difficult to consider.

In bandits, however, the second order probability is a natural interpretation of ambiguity because there is a unique, known second order probability, G. In some applications, it may be difficult to argue that this probability is in fact known, but for the abstract version of the problem that can be tested in the laboratory, a sensible definition of ambiguity can be based on the variance of G.

Definition 3 *An arm F is **more ambiguous than** an arm F' if the variance of G is greater than the variance of G'.*

Therefore, agents who are ambiguity averse do not like the variance in the distribution from which the means of the payoff distributions are drawn. Although this may not be intuitively distinct from variance in the subjective payoff distribution (especially if, as a good economist, you reduce compound lotteries), ambiguity aversion and risk aversion are only weakly correlated within individuals (Hogarth and Einhorn, 1990).

4.1.1 Models of Ambiguity Aversion

Models of ambiguity aversion fall into three categories, those which leverage unique second order probabilities (and relax reduction of compound lotteries), those which allow multiple probabilities, and those which rely on nonadditive probabilities. Non-additive probability models (e.g., Schmeidler, 1989) do not require that the subjective probabilities of an event in a set which will occur with objective probability one sum to one. Therefore, if an outcome will be either A or B, then $P(A) + P(B)$ does not have to equal one. Therefore, ambiguity aversion is represented because, when computing expected payoffs using the subjective probabilities, not all the probability weight will be represented. The remaining probability, $1 - P(A) - P(B)$, is a measure of faith in the evidence on which agents' beliefs are based.

Another way to think about ambiguity is to consider independently the set of possible payoff distributions without reference to a second order probability which generates them. There are a variety of ways that agents may use these multiple probabilities to decide among actions. Many of these models suggest agents use a

minimax decision rule (e.g., Gilboa and Schmeidler, 1989). Ambiguity aversion arises in these models as agents may have multiple probabilities and pessimistically evaluate each lottery using the probability which generates the lowest expected utility.

Given that bandits provide a known, unique second order probability, and the definition of ambiguity is based on a second order probability, the most natural set of models to apply are those using second order probabilities. These models relax the assumption of reduction of compound lotteries implied by subjective expected utility theories by applying a nonlinear weighting function to transform the second order probability distribution before computing an expectation. When the nonlinear weighting rule moves decision weight from high values to low values, agents will be ambiguity averse.

To formalize the notion of second order probabilities, let Q be a distribution with a parameter θ, where $Q(\theta)$ is the probability the agent will receive some prize X. Let θ have a second order distribution G, and let $\bar{\theta}$ be the expected value of θ.

Ambiguity attitudes are represented by a decision weighting function $\omega(X)$, which gives the expected utility from winning the prize X, adjusting for the ambiguity associated with selecting an ambiguous lottery. Using this, ambiguity aversion can be formally defined.

Definition 4 *An agent is ambiguity averse if her decision weighting function* $\omega(X)$ *has the property that* $E[\omega(X)|F] \leq E[X|F]$.

Formally, ambiguity aversion occurs when the value of a choice given its ambiguity is less than its expected value. Typically, $\omega(X)$ is decreasing in the variance of G.

Several forms for the function $\omega(X)$ have been proposed. Segal (1987) proposes some restrictions on the form of $\omega(X)$ which lead to ambiguity aversion. This chapter focuses on a specific function used by Kahn and Sarin (1988).

$$\omega(X) = u(x)\bar{\theta} + u(x)\int_{\theta=0}^{1}(\theta - \bar{\theta})e^{[-\xi(\theta-\bar{\theta})]/\sigma}dG(\theta)d\theta \tag{4.1}$$

where $\sigma = \sqrt{\int_{\theta=0}^{1}(\theta-\bar{\theta})^2 dG(\theta)d\theta}$ is that standard deviation of the second order probability distribution, and $\bar{\theta}$ is the expected value of θ.

If $\xi \neq 0$, then $(\theta - \bar{\theta})$ is the scale of the impact of the ambiguity. If $\xi > 0$, the second order probabilities are adjusted by underweighting the chance of higher than average θs and overweighting the probabilities of lower than average θs, leading to ambiguity aversion. A negative ξ does the opposite, representing ambiguity preference. Since it captures attitudes toward ambiguity, ξ is a characteristic of the individual, and therefore a primitive of the model.

The decision weighting function can be interpreted as an expectation, where $\bar{\theta}u(x)$ is the expected utility of the choice and the $u(x)$ times the integral is the psychological cost (or reward) associated with making an ambiguous choice. This interpretation requires that the cost be incurred when the choice is made, so it is subtracted from any outcome realized.

4.2 The Gittins Index of Ambiguity Averse Agents

Ambiguity aversion may play an intuitive role in bandit problems. Agents may dislike trying alternatives about which they have less information because they do not know the mean of the payoff distribution; the larger the variance of possible means, the more they dislike unknown alternatives. This attitude could lead to underexperimentation, as agents avoid ambiguous alternatives, and lower than optimal Gittins indexes.

This section formalizes this intuition and develops some theory on how ambiguity aversion affects behavior in bandits. It proves that a Gittins index exists, and that a two-armed bandit with one arm known is a stopping problem for the ambiguity averse agent. These results form the foundation for extending Kahn and Sarin's model of ambiguity aversion to Bernoulli arms with beta priors.

4.2.1 Bayes Rule and Reduction of Compound Lotteries

Extending ambiguity aversion to bandit problems requires some interpretation because these models have, in the past, applied only to choices among compound lotteries. Models which leverage second order probabilities assume that, in computing

their expected payoffs, agents do not use Bayes rule to reduce compound lotteries; they use some other function which expresses their ambiguity aversion.

In bandits, however, ambiguity neutral agents apply Bayes rule not only to reduce compound lotteries to compute their expected payoff, but also to update their priors over the parameters of the payoff distribution and to compute the probabilities of continuations. How to best extend models of choice among compound lotteries to bandits depends on whether ambiguity aversion is the product of a fundamental problem in applying Bayes rule, or whether it is some other phenomenon which is well modeled by using some substitute for Bayes rule.

I am not aware of any evidence on either side of this question. However, for purposes of this paper, I will assume that agents understand and use Bayes rule to update their prior beliefs and to compute continuation probabilities. Therefore, it is only the act of computing an expected value, considering payoffs themselves, in the context of ambiguity which leads to non-Bayesian behavior.

To emphasize this distinction, I will use $E[\omega(X)|F]$ to indicate an expected payoff adjusted for ambiguity. Probabilities will be denoted $P(X = 1|F)$, and are not affected by the agents' ambiguity attitudes. To indicate that a given arm is being evaluated by an ambiguity averse agent, its payoff distribution will be written F_ω. However, because probabilities are unaffected by ambiguity attitudes this does not correspond to a different distribution. Further, because Bayes rule is applied properly in updating prior beliefs, $(X)F_\omega$ reflects that F has been updated to reflect X prior to being evaluated under the ambiguity attitude represented in $\omega(\cdot)$.

4.2.2 Existence of a Dynamic Allocation Index

Before developing a specific model of ambiguity aversion in bandits, I prove that a Gittins index exists for ambiguity averse agents.

Theorem 7 *For each nonincreasing discount sequence A with $A \neq 0$ and each distribution F on $\mathcal{D}$, there exists a unique function $\Lambda(F_\omega, A)$ such that the F arm is optimal initially in the $(F_\omega, \lambda; A)$ bandit if and only if $\lambda \leq \Lambda(F_\omega, A)$ and the λ arm is*

optimal initially if and only if $\lambda \geq \Lambda(F_\omega, A)$.

This proof follows the proof of existence of a Gittins index for hyperbolic discounters. First, it is necessary to prove some preliminary results about $V(F_\omega, \lambda; A)$ and $\Delta(F_\omega, \lambda; A)$.

Lemma 3 *For all* F *and for all* A, $V(F_\omega, \lambda; A)$ *is continuous and nondecreasing in* λ.

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.

Proof: Suppose $\lambda^* > \lambda$ and σ is optimal in the $(F_\omega, \lambda; A)$ bandit. Suppose σ is followed in the $(F_\omega, \lambda^*; A)$ bandit. The only change compared with $(F_\omega, \lambda; A)$ is when arm 2 is selected and λ^* is received instead of λ.

Therefore,

$$\begin{aligned} V(F_\omega, \lambda; A) &= W(F_\omega, \lambda; A; \sigma) && (4.2) \\ &\leq W(F_\omega, \lambda^*; A; \sigma) = V(F_\omega, \lambda^*; A) && (4.3) \end{aligned}$$

Therefore, $V(F_\omega, \lambda; A)$ is nondecreasing in λ.

For continuity, let σ^* be optimal for the $(F_\omega, \lambda^*; A)$. Then the only difference between $W(F_\omega, \lambda; A; \sigma^*)$ and $W(F_\omega, \lambda^*; A; \sigma^*)$ is the result of arm 2 when it is chosen.

$$\begin{aligned} V(F_\omega, \lambda^*; A) &= W(F_\omega, \lambda^*; A) && (4.4) \\ &\leq W(F_\omega, \lambda; A; \sigma^*) + (\lambda^* - \lambda)\gamma_1 && (4.5) \\ &\leq V(F_\omega, \lambda; A) + (\lambda^* - \lambda)\gamma_1 && (4.6) \end{aligned}$$

Since Equation 4.3 proved $V(F_\omega, \lambda; A) \leq V(F_\omega, \lambda^*; A)$, this implies that this relationship must approach equality as $\lambda^* \to \lambda$. Therefore, $V(\cdot, \lambda; A)$ is continuous in λ. □

In order to show a dynamic allocation index exists, I also need a result showing how the size of the error made by choosing the the wrong arm varies with λ. Define the function $\Delta(F_\omega, \lambda; A)$ 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;

$$\Delta(F_\omega, \lambda; A) = V^F(F_\omega, \lambda; A) - V^\lambda(F_\omega, \lambda; A). \tag{4.7}$$

The absolute value of 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 Theorem 7.

Lemma 4 *For any F on $\mathcal{D}$, $\Delta(F_\omega, \lambda; A)$ is decreasing in λ when A is nonincreasing with $A \neq 0$.*

Proof: Fix $\lambda^* > \lambda$.

This proof proceeds in three parts. Part (i) derives an expression for $\Delta(F_\omega, \lambda^*; A) - \Delta(F_\omega, \lambda; A)$. Part (ii) performs a finite induction on the horizon to establish that $\Delta(F_\omega, \lambda^*; A) - \Delta(F_\omega, \lambda; A)$ is negative. 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

$$V^F(F_\omega, \lambda; A) = \alpha_1 E[\omega(X)|F] + E[V((X)F_\omega, \lambda; A^{(1)})]. \tag{4.8}$$

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

$$V^\lambda(F_\omega, \lambda; A) = \alpha_1 \lambda + V(F_\omega, \lambda; A^{(1)}). \tag{4.9}$$

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

$$V^F(F_\omega, \lambda; A) = V(F_\omega, \lambda; A) - \Delta^-(F_\omega, \lambda; A) \tag{4.10}$$

$$V^\lambda(F_\omega, \lambda; A) = V(F_\omega, \lambda; A) - \Delta^+(F_\omega, \lambda; A). \tag{4.11}$$

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

Using these definitions, substitute for $V((X)F_\omega, \lambda; A)$ and $V(F_\omega, \lambda; A)$ in Equations 4.8 and 4.9 above. This gives

$$V^F(F_\omega, \lambda; A) = \alpha_1 E[\omega(X)|F] + E[V^\lambda((X)F_\omega, \lambda; A^{(1)}) + \Delta^+((X)F_\omega, \lambda; A^{(1)})] \tag{4.12}$$

$$V^\lambda(F_\omega, \lambda; A) = \alpha_1 \lambda + V^F(F_\omega, \lambda; A^{(1)}) + \Delta^-(F_\omega, \lambda; A^{(1)}). \tag{4.13}$$

These expressions can then be used to compute $\Delta(F_\omega, \lambda; A)$. The first two terms in Equation 4.12 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 3.14 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 4.13 from those in Equation 4.12 gives $(\alpha_1 - \alpha_2)[E[\omega(X)|F] - \lambda]$. This gives

$$\Delta(F_\omega, \lambda; A) = (\alpha_1 - \alpha_2)[E[\omega(X)|F] - \lambda] + E[\Delta^+((X)F_\omega, \lambda; A^{(1)})] - \Delta^-(F_\omega, \lambda; A^{(1)}). \tag{4.14}$$

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

$$\begin{aligned}\Delta(F_\omega, \lambda^*; A) - \Delta(F_\omega, \lambda; A) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^+((X)F_\omega, \lambda^*; A^{(1)}) - \Delta^+((X)F_\omega, \lambda; A^{(1)})] + \\ \Delta^-(F_\omega, \lambda; A^{(1)}) - \Delta^-(F_\omega, \lambda^*; A^{(1)}).\end{aligned} \tag{4.15}$$

(ii) Proving the lemma requires that the difference in Equation 4.15 be negative. 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, $\Delta(F_\omega, \lambda^*; A_1) - \Delta(F_\omega, \lambda; A_1)$ is negative implies

$$\begin{aligned} E[\omega(X)|F] - \lambda^* &< E[\omega(X)|F] - \lambda \\ \lambda^* &> \lambda \end{aligned} \tag{4.16}$$

which is true by assumption.

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

Equation 4.15 can be rewritten with the truncated discount sequence

$$\begin{aligned} \Delta(F_\omega, \lambda^*; A_{n+1}) - \Delta(F_\omega, \lambda; A_{n+1}) = (\alpha_1 - \alpha_2)[\lambda - \lambda^*] + \\ E[\Delta^+((X)F_\omega, \lambda^*; A^{(1)}_{n+1}) - \Delta^+((X)F_\omega, \lambda; A^{(1)}_{n+1})] + \\ \Delta^-(F_\omega, \lambda; A^{(1)}_{n+1}) - \Delta^-(F_\omega, \lambda^*; A^{(1)}_{n+1}). \end{aligned} \tag{4.17}$$

The first term on the right-hand side of Equation 4.17 is nonpositive because $\lambda^* > \lambda$ by assumption and $\alpha_1 \geq \alpha_2$ by hypothesis.

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

$$\begin{aligned} &E[\Delta^+((X)F_\omega, \lambda^*; A^{(1)}_{n+1}) - \Delta^+((X)F_\omega, \lambda; A^{(1)}_{n+1})] \\ &= E[\max[0, \Delta((X)F_\omega, \lambda^*; A_n)] - \max[0, \Delta((X)F_\omega, \lambda; A_n)]]. \end{aligned} \tag{4.18}$$

The induction hypothesis gives that $\Delta(F_\omega, \lambda^*; A_n) < \Delta(F_\omega, \lambda; A_n)$ for all F, in particular σF. Therefore, the second term in Equation 4.18 is always larger than the

first, implying that the second term in Equation 4.17 is negative.

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

$$\begin{aligned} &\Delta^-(F_\omega, \lambda; A^{(1)}_{n+1}) - \Delta^-(F_\omega, \lambda^*; A^{(1)}_{n+1}) \\ &= \max[0, -\Delta(F_\omega, \lambda; A_n)] - \max[0, -\Delta(F_\omega, \lambda^*; A_n)]. \end{aligned} \tag{4.19}$$

The induction hypothesis gives that $\Delta(F_\omega, \lambda^*; A_n) \leq \Delta(F_\omega, \lambda; A_n)$. Therefore, the second term in Equation 4.19 is always larger than the first, implying that the third term in Equation 4.17 is negative.

Since the first term in Equation 4.17 is nonpositive and the other two are strictly negative, the difference Equation 4.15 is negative for every finite horizon. Now we let the horizon go to infinity to show it is negative 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

$$\Delta(F_\omega, \lambda^*; A_\infty) < \Delta(F_\omega, \lambda; A_\infty). \tag{4.20}$$

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

Given this result, the existence of a dynamic allocation index is easy to prove.

Proof of Theorem 7: This proof begins by defining $\Lambda(F_\omega, A) = \inf\{\lambda \in \mathcal{D} :$ the λ arm is optimal for the $(F_\omega, \lambda; A)$ bandit $\}$. Then I show that this definition implies that F is uniquely optimal if $\lambda < \Lambda(F_\omega, A)$. Then I use Lemma 4 to show that λ is uniquely optimal if $\lambda > \Lambda(F_\omega, A)$. Indifference at $\lambda = \Lambda(F_\omega, A)$ then follows from the continuity of V.

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

$$V^F(F_\omega, \lambda; A) > V^\lambda(F_\omega, \lambda; A) \tag{4.21}$$

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

The case where $\lambda > \Lambda(F_\omega, A)$ is a little harder because there may be values of λ above $\Lambda(F_\omega, A)$ where F is optimal. However, the fact that $\Delta(F_\omega, \lambda; A)$ is strictly decreasing in λ, as shown in Lemma 4, proves that this cannot be. Therefore,

$$V^F(F_\omega, \lambda; A) < V^\lambda(F_\omega, \lambda; A) \tag{4.22}$$

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

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

$$V^F(F_\omega, \Lambda(F_\omega, A); A) = V^\lambda(F_\omega, \Lambda(F_\omega, A); A), \tag{4.23}$$

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

4.2.3 The Stopping Property

The index will be much easier to compute and discuss theoretically if ambiguity averse agents' strategies satisfy the stopping property.

Theorem 8 *If A is nonincreasing, then for every $(F_\omega, \lambda; A)$ bandit, there is an optimal strategy for which every selection of λ is followed by another selection of λ.*

First, I need a result which indicates when λ will never be optimal.

Lemma 5 *If $E[\omega(X)|F] \geq \lambda$, then the F arm is optimal for any A.*

Proof: Suppose that σ is an optimal strategy for the $(F_\omega, \lambda; A^{(1)})$ bandit. Let σ^* be a strategy which indicates the F arm initially and then follows σ, ignoring the initial realization from F. Then

$$\begin{aligned} W(F_\omega, \lambda; A; \sigma^*) &= \alpha_1 E[\omega(X)|F] + V(F_\omega, \lambda; A^{(1)}) && (4.24) \\ &\geq \alpha_1 \lambda + V(F_\omega, \lambda; A^{(1)}) = W(F_\omega, \lambda; A; \sigma_2) && (4.25) \end{aligned}$$

where σ_2 is a strategy which chooses λ initially and then proceeds optimally. Since there is a strategy which starts with the F arm and is least as good as the optimal strategy which starts with the λ arm, the F arm is optimal. □

Proof of Theorem 8: Let A_n denote a nonincreasing discount sequence with horizon n. The proof is by induction on the horizon.

If $n = 1$, then the proposition is trivially true since there is no further selection of either arm.

Suppose $n \geq 2$ and for every $(F_\omega, \lambda; A_{n-1})$ bandit, there is an optimal strategy for which every selection of λ is followed by another selection of λ. Then $A_n^{(1)} \in A_{n-1}$.

Assume it is optimal to select F initially. Then the inductive hypothesis shows that there exists a continuation which never switches back to F after its first selection of λ.

On the other hand, if it is optimal to select λ initially, the inductive hypothesis applies trivially unless there is an optimal strategy σ^* which indicates a selection of λ, then selects F at stages $2 \ldots N$, and then λ thereafter. This might be the case if the agent were indifferent at time 1 and returned to indifference after several selection of F. I will now show such a strategy cannot be better than a strategy which never switches back to F.

Since the value of λ is known, we can assume that σ^* does not depend on the initial observation from λ. So for each m, $\{N > m\}$ is measurable with respect to the σ-field generated on the outcomes $(X_2, \ldots, X_m)$.

Lemma 5 implies that if the sequence of outcomes through time m contains s successes and $f = m - s - 1$ failures while following σ^*, then

$$N = m \Rightarrow E[X|\sigma^s \phi^f F_\omega] < \lambda \tag{4.26}$$

This condition says that if λ is going to be selected in the current period, then the sequence of successes and failures must be such that it is no longer optimal to select F. This bound comes from Lemma 5.

Now I will show that there is a strategy σ which starts with F and is at least as good as σ^*. Let σ select F initially and then imitate σ^* by selecting the arm prescribed by σ^* one period earlier.

$$W(F_\omega, \lambda; A; \sigma^*) = E_{\sigma^*}\left[\alpha_1\lambda + \sum_{m=2}^{N} \omega(X_m)\alpha_m + \lambda \sum_{m=N+1}^{\infty} \alpha_m | F\right] \tag{4.27}$$

which must be at least $\gamma_1\lambda$ since σ^* is optimal. Therefore,

$$\sum_{m=2}^{\infty} E_{\sigma^*}\left[(\omega(X_m) - \lambda)I\{N \geq m\}|F\right]\alpha_m = E_{\sigma^*}\left[\sum_{m=2}^{N}(\omega(X_m) - \lambda)\alpha_m|F\right] \geq 0 \tag{4.28}$$

The value under σ is given by

$$W(F_\omega, \lambda; A; \sigma) = E_\sigma\left[\sum_{m=2}^{N} \omega(X_m)\alpha_{m-1} + \lambda \sum_{m=N+1}^{\infty} \alpha_{m-1}|F\right] \tag{4.29}$$

The stopping strategy σ is better if the difference in these worths is positive.

$$W(F_\omega, \lambda; A; \sigma) - W(F_\omega, \lambda; A; \sigma^*) = \sum_{m=2}^{\infty} E_{\sigma^*}\left[(\omega(X_m) - \lambda)I\{N \geq m\}|F\right](\alpha_{m-1} - \alpha_m) \tag{4.30}$$

The term $(\alpha_{m-1} - \alpha_m)$ is weakly positive because A_n is nonincreasing. The expectation weakly is positive because $E[\omega(X_m)|F] \leq \lambda$ by Lemma 5.

Therefore, there is a strategy which satisfies the stopping property and which is at least as good as the optimal strategy.□

4.3 Bandits with Kahn-Sarin Ambiguity Averse Agents

This section establishes properties of the behavior of ambiguity averse agents in bandit problems. The analysis here is restricted to the case of a Bernoulli arm whose parameter has a known beta(α, β) distribution.

The primary concern of this section is that adding the dynamic element of probability updating to the ambiguous choice problem does not affect the way the model represents ambiguity aversion. It is not obvious from Kahn and Sarin's paper that, when an agent acquires more information about an ambiguous alternative, $\xi > 0$ will still lead to an $\omega(\cdot)$ which is ambiguity averse. As it turns out, we do not need to be concerned that we will observe some sequence of successes and failures for which the Kahn-Sarin transformation will result in behavior other than ambiguity aversion.

Theorem 9 *Let F be the distribution of the parameter of a Bernoulli arm with a beta(α, β) prior. If ξ is positive, then the Kahn-Sarin agent will be ambiguity averse for all informative priors ($\alpha > 0$ and $\beta > 0$).*

The proof of this theorem reduces the problem to a difference between two Kummer's functions, or confluent hypergeometric functions of the first kind. (Spanier and Oldham, 1987; Abramowitz and Stegun, 1972) This function is

$$ {}_1F_1(a, b, z) = \frac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)} \int_0^1 e^{zt} t^{a-1} (1-t)^{b-a-1} dt. \tag{4.31} $$

Kummer's function has many uses in theoretical physics, and it provides one class of solutions to Kummer's confluent hypergeometric differential equation, $zy'' + (b - z)y' - ay = 0$ with initial conditions ${}_1F_1(a, b, 0) = 1$ and $\frac{\partial}{\partial z} {}_1F_1(a, b, z)|_{z=0} = a/b$. There is no obvious relationship between its physical interpretation and its appearance in a model of ambiguity aversion; it is used here because existing results about Kummer's function simplify proof of Theorem 9.

Before proving Theorem 9, it is necessary to prove a property of Kummer's function.

Lemma 6 *Kummer's confluent hypergeometric function has the property that ${}_1F_1(a+1, b+1, z) - {}_1F_1(a, b, z)$ is strictly negative for all $b > a > 0$ and $z < 0$.*

Proof: Kummer's first theorem allows $z < 0$ to be transformed into $z > 0$. Then Kummer's function can be related to an Euler beta function, which provides a closed-form solution to the definite integral in ${}_1F_1(\cdot, \cdot, z)$. The closed-form solution can

be manipulated to demonstrate that each term in the summation in the Euler beta function is strictly negative.

To keep track of the sign of z, let z_- denote $z < 0$, and $z_+ = -z_-$ denote $z > 0$. Kummer's first theorem (Slater, 1960, p. 6) holds that ${}_1F_1(a, b, z) = e^z \; {}_1F_1(b - a, b, -z)$ for all z. Therefore,

$$
\begin{aligned}
{}_1F_1(a, b, z_-) &= e^{z_-} \; {}_1F_1(b - a, b, z_+) && (4.32)\\
{}_1F_1(a + 1, b + 1, z_-) &= e^{z_-} \; {}_1F_1(b - a, b + 1, z_+). && (4.33)
\end{aligned}
$$

The difference then reduces to

$$
\frac{\Gamma(b)e^{z_-}}{\Gamma(b-a)\Gamma(a)} \left[\frac{b}{a} \int_0^1 e^{tz_+} t^{b-a-1}(1-t)^a dt - \int_0^1 e^{tz_+} t^{b-a-1}(1-t)^{a-1} dt \right]. \qquad (4.34)
$$

The exponential term in the integral can be broken into its representation as an infinite sum (Slater, 1960, p. 34), yielding

$$
\begin{aligned}
&\frac{\Gamma(b)e^{z_-}}{\Gamma(b-a)\Gamma(a)} \left[\frac{b}{a} \int_0^1 \sum_{n=0}^{\infty} \frac{z_+^n}{n!} t^{b-a-1+n}(1-t)^a dt \right.\\
&\qquad \left. - \int_0^1 \sum_{n=0}^{\infty} \frac{z_+^n}{n!} t^{b-a-1+n}(1-t)^{a-1} dt \right]. \qquad (4.35)
\end{aligned}
$$

The terms which depend on z_+ can be moved out of the integral, and the difference can be written

$$
\frac{\Gamma(b)e^{z_-}}{\Gamma(b-a)\Gamma(a)} \sum_{n=0}^{\infty} \frac{z_+^n}{n!} \int_0^1 \left(\frac{b}{a}(1-t) - 1 \right) t^{b-a-1+n}(1-t)^{a-1} dt. \qquad (4.36)
$$

The integral is a (difference of) Euler beta functions, and its solution can therefore be represented in terms of gamma functions. The integral equals

$$
\begin{aligned}
&= \Gamma(a) \frac{(b-a)\Gamma(b+1+n)\Gamma(b-a+n) - b\Gamma(b+n)\Gamma(b-a+1+n)}{a\Gamma(b+n)\Gamma(b+1+n)} && (4.37)\\
&= \frac{\Gamma(a)\Gamma(b-a+n)}{a(b+n)\Gamma(b+n)} \left[(b-a)(b+n) - b(b-a+n) \right] && (4.38)
\end{aligned}
$$

$$= -an\frac{\Gamma(a)\Gamma(b-a+n)}{a(b+n)\Gamma(b+n)}. \tag{4.39}$$

The entire difference can be represented as

$$-\frac{\Gamma(b)e^{z-}}{\Gamma(b-a)\Gamma(a)}\sum_{n=0}^{\infty}\frac{z_+^n}{n!}\frac{n\Gamma(a)\Gamma(b-a+n)}{(b+n)\Gamma(b+n)}. \tag{4.40}$$

Since $b > 0$, $a > 0$, $b - a > 0$ and $n > 0$, every gamma function in this expression is strictly positive. Similarly, $z_+ > 0$, so every power of z_+ is strictly positive also. Therefore, every term in the infinite sum is strictly positive. Additionally, $e^{z-} > 0$ is strictly positive. Therefore, the entire difference is negative for any $b > a > 0$. □

The proof that $z > 0$ leads to a positive difference (and ultimately ambiguity-seeking behavior) follows along similar lines. However, the application of Kummer's First Theorem is not necessary because all powers of z are positive, without needing to change its sign.

I can now prove the main result.

Proof of Theorem 9: This proof manipulates the Kahn-Sarin model of ambiguity aversion with a beta distribution (the conjugate prior for the Bernoulli distribution). The ambiguity term in the decision weight expression is shown to be proportional to the difference between two confluent hypergeometric functions of the first kind. Lemma 6 proves this difference is strictly negative, implying ambiguity aversion.

The Kahn-Sarin model holds that agents make choices based on a decision weight equal to

$$E[\omega(x)|\alpha,\beta] = \bar{\theta} + \int_0^1 (\theta-\bar{\theta})e^{-\xi(\theta-\bar{\theta})/\sigma}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \tag{4.41}$$

where $\bar{\theta} = \frac{\alpha}{\alpha+\beta}$ is the expected value of x and $\sigma = \sqrt{\frac{\alpha+\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$ is the standard deviation of x.

Ambiguity aversion holds when $E[\omega(x)|\alpha,\beta] < E[x|\alpha,\beta]$. This is equivalent to

showing that the integral is strictly negative. Evaluating it gives

$$\begin{aligned}
&\frac{1}{\Gamma(\alpha)\Gamma(\beta)}\left[\Gamma(\alpha+1)\Gamma(\beta)e^{\xi\bar{\theta}/\sigma}\ {}_1F_1(\alpha+1,\alpha+\beta+1,-\xi/\sigma)\right.\\
&\qquad\left.-\Gamma(\alpha)\Gamma(\beta)\bar{\theta}e^{\xi\bar{\theta}/\sigma}\ {}_1F_1(\alpha,\alpha+\beta,-\xi/\sigma)\right]\\
&=\frac{e^{\xi\bar{\theta}/\sigma}}{\Gamma(\alpha)}\left[\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)}\ {}_1F_1(\alpha+1,\alpha+\beta+1,-\xi/\sigma)\right.\\
&\qquad\left.-\frac{\bar{\theta}\Gamma(\alpha)}{\Gamma(\alpha+\beta)}\ {}_1F_1(\alpha,\alpha+\beta,-\xi/\sigma)\right]. \qquad (4.42)
\end{aligned}$$

Applying the fact that $\Gamma(k+1)=k\Gamma(k)$ on the numerator and denominator of the left-hand term gives

$$\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)}=\frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)}=\bar{\theta}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)}. \qquad (4.43)$$

The integral therefore evaluates to

$$\frac{\bar{\theta}e^{\xi\bar{\theta}/\sigma}}{\Gamma(\alpha+\beta)}\left[\ {}_1F_1(\alpha+1,\alpha+\beta+1,-\xi/\sigma)-\ {}_1F_1(\alpha,\alpha+\beta,-\xi/\sigma)\right]. \qquad (4.44)$$

Since the coefficient is positive under the assumptions of the theorem, the sign of the integral is determined by the sign of the difference between the Kummer's functions. Lemma 6 proves that this difference is strictly negative. Therefore, the value of the integral is strictly negative, and $E[\omega(x)|\alpha,\beta]<E[x|\alpha,\beta]$, which proves the claim. □

Corrolary 1 *Kahn-Sarin ambiguity aversion is preserved under all possible sequences of successes and failures.*

Bayes rule prescribes that the posterior of a beta distribution with Bernoulli observations is given by beta$(\alpha+s,\beta+f)$ after observing s successes and f failures. Since $s\geq 0$ and $f\geq 0$, $\alpha'=\alpha+s>0$ and $\beta'=\beta+f>0$. Therefore, Theorem 9 applies with beta(α',β'). □

4.3.1 The Gittins Index of an Ambiguity Averse Agent

For ambiguity aversion to be a descriptive model of behavior, it should predict the lower-than-optimal Gittins indexes observed.

Theorem 10 *Suppose A is regular with $\alpha_1 > 0$ and F_ω is the ambiguity-aversion adjusted version of F. Then $\Lambda(F, A) \geq \Lambda(F_\omega, A)$.*

Before proving this directly, we need a result about the impact of ambiguity aversion on the value function.

Lemma 7 *Suppose F is a Bernoulli arm and the agent is ambiguity averse. Then*

$$V(F, \lambda; A) \geq V(F_\omega, \lambda; A) \tag{4.45}$$

for all A.

Proof: This proof proceeds by induction on the horizon.

Suppose the horizon of A is zero. Then

$$V(F, \lambda; A_0) = V(F_\omega, \lambda; A_0) = 0 \tag{4.46}$$

because neither arm is ever selected and no reward is received.

For any $n \geq 1$, assume that $V(F, \lambda; A_n) \geq V(F_\omega, \lambda; A_n)$ for any F satisfying the conditions of the proposition and for any A with horizon less than n.

Suppose the horizon of A is n, and τ is an optimal strategy in the $(F_\omega, \lambda; A_n)$ bandit. Let τ' have the same first selection as τ and then proceed optimally in the $(F, \lambda; A_n)$ bandit.

It is sufficient to show that

$$W(F, \lambda; A_n; \tau') \geq V(F_\omega, \lambda; A_n). \tag{4.47}$$

Without loss of generality, assume that the first move under both τ and τ' is arm 1. Then

$$\begin{aligned} W(F,\lambda;A_n;\tau') &= \alpha_1 E[X|F] + P(X=1|F)V(\sigma F,\lambda;A_n^{(1)}) \\ &\quad + P(X=0|F)V(\phi F,\lambda;A_n^{(1)}) \\ V(F_\omega,\lambda;A_n) &= \alpha_1 E[\omega(X)|F] + P(X=1|F)V(\sigma F_\omega,\lambda;A_n^{(1)}) \\ &\quad + P(X=0|F)V(\phi F_\omega,\lambda;A_n^{(1)}). \end{aligned} \tag{4.48}$$

Note that the probabilities $E[X|F]$ in the second equation are not transformed by the ambiguity function; the agent understands the probabilities, but does not like the uncertainty they imply for her payoffs.

$$\begin{aligned} & W(F,\lambda;A_n;\tau') - V(F_\omega,\lambda;A_n) \\ &= \alpha_1 \left[E[X|F] - E[\omega(X)|F_\omega]\right] \\ &\quad + P(X=1|F)\left[V(\sigma F,\lambda;A_n^{(1)}) - V(\sigma F_\omega,\lambda;A_n^{(1)})\right] \\ &\quad + P(X=0|F)\left[V(\phi F,\lambda;A_n^{(1)}) - V(\phi F_\omega,\lambda;A_n^{(1)})\right] \end{aligned}$$

The first term is nonnegative because $\omega()$ represents ambiguity aversion.

The second term is nonnegative because the induction hypothesis applies to any F, in particular σF. The third term is nonnegative because the induction hypothesis applies to any F, in particular ϕF.

Therefore, the difference is nonnegative, so the claim is proven. □

Proof of Theorem 10: This proof is by contradiction.

Suppose $\Lambda(F,A) < \Lambda(F_\omega,A)$. Then arm 1 is optimal initially in the $(F,\Lambda(F,A);A)$ and $(F_\omega,\Lambda(F,A);A)$ bandits. Then

$$\begin{aligned} & V(F,\Lambda(F,A);A) - V(F_\omega,\Lambda(F,A);A) \\ &= \alpha_1 \left[E[X|F] - E[\omega(X)|F]\right] \\ &\quad + P(X=1|F)\left[V(\sigma F,\Lambda(F,A);A^{(1)}) - V(\sigma F_\omega,\Lambda(F,A);A^{(1)})\right] \end{aligned}$$

$$+ P(X = 0|F)\left[V(\phi F, \Lambda(F,A); A^{(1)}) - V(\phi F_\omega, \Lambda(F,A); A^{(1)})\right]. \quad (4.49)$$

The first term is nonnegative by the definition of ambiguity aversion. The second and third terms are each positive by the proposition above.

However, we also know that the strategy using arm 2 at every stage is optimal for the $(F, \Lambda(F,A); A)$ bandit. Therefore,

$$\begin{aligned} V(F, \Lambda(F,A); A) - V(F_\omega, \Lambda(F,A); A) &= \gamma_1 \Lambda(F,A) - V(F_\omega, \Lambda(F,A); A) \quad (4.50) \\ &\leq 0. \end{aligned}$$

The inequality follows because $\Lambda(F,A) < \Lambda(F_\omega, A)$ implies that arm 2 cannot be the optimal choice in the first period of the $(F_\omega, \Lambda(F,A); A)$ bandit. Therefore, the value of the $(F_\omega, \Lambda(F,A); A)$ bandit must be strictly greater than $\gamma_1 \Lambda(F,A)$ (arrived at through some strategy which selects arm 1 initially). However, Equation 4.50 being negative contradicts Equation 4.49 being positive. Therefore, it cannot be that $\Lambda(F,A) < \Lambda(F_\omega, A)$. □

4.3.2 Willingness to Pay for Information About the True Average Payoff of an Ambiguous Arm

Theorem 11 *If A is regular, an ambiguity averse agent will pay more to learn the value of the parameter θ than an optimal agent.*

Proof: This proof calculates the value of learning the value of θ and demonstrates that it is increasing in λ. The result follows from Theorem 10 which demonstrates that $\Lambda(F,A) \geq \Lambda(F_\omega, A)$.

Because A is regular, the value of playing an uncertain bandit is given by $\gamma_1 \Lambda(F,A)$.

If agents knew the value of θ, they would select the arm yielding $\Lambda(F,A)$ if $\theta < \Lambda(F,A)$, and the arm yielding θ otherwise. This yields a payoff of

$$\gamma_1 \left[\int_0^{\Lambda(F,A)} \Lambda(F,A) dF(\theta) d\theta + \int_{\Lambda(F,A)}^1 \theta dF(\theta) d\theta\right]. \quad (4.51)$$

The value of knowing θ, or the amount the agent would be willing to pay to learn θ, is given by

$$\begin{aligned}
&\gamma_1 \left[\int_0^{\Lambda(F,A)} \Lambda(F,A) dF(\theta) d\theta + \int_{\Lambda(F,A)}^1 \theta dF(\theta) d\theta \right] - \gamma_1 \Lambda(F,A) \\
&= \gamma_1 \left[\int_0^{\Lambda(F,A)} \Lambda(F,A) dF(\theta) d\theta + \int_{\Lambda(F,A)}^1 \theta dF(\theta) d\theta \right] \\
&\quad - \gamma_1 \left[\int_0^{\Lambda(F,A)} \Lambda(F,A) dF(\theta) d\theta + \int_{\Lambda(F,A)}^1 \Lambda(F,A) dF(\theta) d\theta \right] \\
&= \gamma_1 \int_{\Lambda(F,A)}^1 (\theta - \Lambda(F,A)) dF(\theta) d\theta.
\end{aligned}$$

This expression is positive since $(\theta - \Lambda(F,A)) \geq 0$ at every point in the range of integration. It is zero only if $\Lambda(F,A) = 1$.

It remains to show that this value is decreasing in λ. Now suppose $\lambda^* > \lambda$. This will show that changing from λ^* to λ results in an increase in the value of learning θ.

At λ^*, the value of learning θ is

$$\gamma_1 \int_{\lambda^*}^1 (\theta - \lambda^*) dF(\theta) d\theta. \tag{4.52}$$

A smaller index has two effects on this quantity. First, it extends the range of the integral to the range of $\lambda \leq \theta < \lambda^*$. Second, it increases the integrand by increasing the difference. The value of learning θ under λ can be written

$$\gamma_1 \left[\int_{\lambda}^{\lambda^*} (\theta - \lambda) dF(\theta) d\theta + \int_{\lambda^*}^1 \left[(\theta - \lambda^*) + (\lambda^* - \lambda) \right] dF(\theta) d\theta \right]. \tag{4.53}$$

Both of the new terms are positive, indicating that the value is decreasing in λ.

Since $\Lambda(F,A) \geq \Lambda(F_\omega, A)$, the ambiguity averse agent will be willing to pay more for information about the value of θ. □

The result of this theorem provides a surprising contrast to Theorem 10. Because agents are willing to pay more than optimal to learn about the value of θ, it appears that they overvalue information. On the other hand, because their Gittins index is lower than optimal, ambiguity averse agents appear to undervalue information.

This paradox does not arise in the other models studied. Ambiguity averse agents value the counterfactual universe where they know θ without ambiguity aversion; agents who are hyperbolic discounters, or who cannot properly solve the dynamic programming problem, will bring these suboptimalities to the counterfactual calculation. Therefore, this surprising prediction of ambiguity aversion is excellent grounds for testing the model.

A slightly different result is needed to test this prediction in an experiment. Because I do not know the agents Gittins indexes, I cannot set the value of the second arm at the value that makes the agent indifferent. However, Theorem 11 can be generalized.

Theorem 12 *An ambiguity averse agent will pay more than an ambiguity neutral agent for information about the true value of θ in the $(F, \lambda; A)$ for any λ. If $\lambda < \Lambda(F, A)$, the ambiguity averse agent will pay strictly more.*

Proof: If the agent knows θ, then her expected payoff will be $\gamma_1 \max[\theta, \lambda]$ because the agent will simply select whichever arm gives the higher payoff.

If the agent does not know θ, she computes a Gittins index $\Lambda(F., A)$ (where $F.$ is either F or F_ω). Then she expects to receive $V(F., \lambda; A)$ if $\lambda < \Lambda(F., A)$ and $\gamma_1 \lambda$ if $\lambda \geq \Lambda(F., A)$.

Given that $\max[\theta, \lambda]$ can be written as in Equation 4.51, the agents should be willing to pay

$$\gamma_1 \left[\int_0^\lambda \lambda dF(\theta) d\theta + \int_\lambda^1 \theta dF(\theta) d\theta - \lambda \right] \tag{4.54}$$

if $\lambda > \Lambda(\cdot, A)$, and

$$\gamma_1 \left[\int_0^\lambda \lambda dF(\theta) d\theta + \int_\lambda^1 \theta dF(\theta) d\theta \right] - V(F, \lambda; A) \tag{4.55}$$

if $\lambda < \Lambda(\cdot, A)$.

There are three cases to consider.

(i) If $\lambda \leq \Lambda(F_\omega, A) < \Lambda(F, A)$, then both an ambiguity averse and an ambiguity neutral agent will select F in the first period, and Equation 4.55 applies. Lemma

7 shows $V(F_\omega, \lambda; A) < V(F, \lambda; A)$, so Equation 4.55 will be larger for the ambiguity averse agent.

(ii) If $\Lambda(F_\omega, A) < \Lambda(F, A) \leq \lambda$, then it is optimal for both agents to select λ in the first period (and indefinitely), so they will both determine the value of learning θ using Equation 4.54. This value does not depend on attitude toward ambiguity, so both agents will pay the same amount.

(iii) If $\Lambda(F_\omega, A) \leq \lambda < \Lambda(F, A)$, then the ambiguity averse agent will select λ in the first period and determine the value of learning θ using Equation 4.54, but the ambiguity neutral agent will select F in the first period and determine the value of learning θ using Equation 4.55.

For any F and any value of λ, Equation 4.54 is larger than Equation 4.55. Subtracting Equation 4.55 from Equation 4.54 yields

$$V(V, \lambda; A) - \gamma_1 \lambda > 0. \tag{4.56}$$

Positivity follows because the fact that an agent is using Equation 4.55 implies F is the optimal first period choice, and for this to be the case, it must allow the agent to do better than $\gamma_1 \lambda$. Therefore, Equation 4.54 is larger than Equation 4.55, and the ambiguity averse agent will pay more to learn θ. □

Note that the ambiguity averse agent will pay strictly more in cases (i) and (iii), and exactly the same in case (ii).

4.4 Alternative Explanations

In addition to generating surprising and testable prediction for ambiguity aversion, the idea of willingness to pay provides a useful testbed for other theories which may explain initial underexperimentation. The hyperbolic discounting model considered in Chapter 3 makes a testable prediction about willingness to pay presented in Section 4.5.6. This section develops the framework within which two other models, risk aversion and quantal response strategies, also make predictions in the willingness to

pay treatment.

4.4.1 Risk Aversion

Previous sections have controlled for the effect of payoff variance by holding constant the mean-conditional distribution of payoffs, Q, across arms. It is likely, however, that risk averse agents are responsive to their subjective payoff distributions, which are affected by both the prior and payoff distributions. The willingness to pay treatment allows a direct test of aversion to the variance of F.

A risk averse agent who does not discount and who is ambiguity neutral will treat the finite horizon bandit as a T-stage compound lottery. Using Bayes rule, this lottery can be reduced to a single stage lottery over the possible states of final wealth. For sufficiently unfavorable priors over the mean of the unknown arm, the known arm is prescribed in each period. The expected utility of this strategy is given by

$$\sum_{s=0}^{T} \binom{T}{s} \lambda^s (1-\lambda)^{T-s} U(\omega_0 + s) \tag{4.57}$$

where ω_0 is the initial wealth level.

However, an agent who knows the value of both arms will select the arm with the higher mean in each period.

$$G(\lambda) \sum_{s=0}^{T} \binom{T}{s} \lambda^s (1-\lambda)^{T-s} U(\omega_0 + s) + \int_{\lambda}^{1} \sum_{s=0}^{T} \binom{T}{s} \theta^s (1-\theta)^{T-s} U(\omega_0 + s) dG(\theta) d\theta \tag{4.58}$$

Equation 4.58 is larger than Equation 4.57 because it is a convex combination of Equation 4.57 and something larger. Therefore, the risk averse agent should be willing to pay some amount to learn the true mean of the unknown arm. The agent's maximum willingness to pay is $\omega_0 - \omega$, where ω satisfies the following condition:

$$\sum_{s=0}^{T} \binom{T}{s} \lambda^s (1-\lambda)^{T-s} U(\omega_0 + s)$$

$$= \sum_{s=0}^{T} \binom{T}{s} \left[G(\lambda)\lambda^s(1-\lambda)^{T-s} + \int_{\lambda}^{1} \theta^s(1-\theta)^{T-s} dG(\theta)d\theta \right] U(\omega+s). \quad (4.59)$$

If the utility function is linear $\omega_0 - \omega$ is exactly the optimal, risk neutral value.

It is not obvious how the willingness to pay of the risk averse agent will differ from that of the risk neutral agent. In fact, this questions is easiest to address numerically. Section 4.5.6 demonstrates that, under certain circumstances, risk aversion predicts lower than optimal willingness to pay.

4.4.2 Quantal Response

Another model which is intuitively consistent with previous data is McKelvey and Palfrey's (1995, 1998) idea of quantal response. They suggest that players in games may make errors in strategy selection whose likelihoods are proportional to the differences in payoffs arising from each strategy. Furthermore, if a player anticipates these errors on the part of other players in the game, optimal strategies may change. This model seems to explain deviations from extreme point-predictions in the laboratory, like passing in the centipede game and making a large offer in the ultimatum game.

Quantal response may explain the lower than optimal initial Gittins indexes, as well as the small amount of overexperimentation observed in later periods. Initially, Gittins indexes might be smaller than optimal because probabilistic errors will prevent agents from taking full advantage of the information they acquire: they will be able to act optimally most of the time, but sometimes they will select the wrong arm, even if they have acquired the correct information to that point. Conversely, Gittins indexes may be small because agents will sometimes experiment "by mistake," reducing the need to do so purposefully. In later periods, agents will select arms optimal agents would not because of a probabilistic error. Therefore, quantal response agents will appear to experiment too much.

Adapting quantal response to bandit problems requires a natural extension of the agent form of the extensive form game model outlined in McKelvey and Palfrey (1998). In that model, the agent at each node assumes that agents at all other nodes

are going to make errors according to a particular error function. For bandit problems, these agents are the agent herself at every possible future state. Following McKelvey and Palfrey, this analysis focuses on a logistic error function, which means that the probability of selecting the known arm with value λ in a state defined by beliefs F is

$$Pr(\text{choice} = \lambda|\lambda, F; A) = \frac{\exp[\eta\lambda]}{\exp[\eta\lambda] + \exp[\eta\Lambda_q(F, A)]} \tag{4.60}$$

where $\Lambda_q(F, A)$ is the optimal index function for a quantal response agent. The parameter η is the precision, or error rate, of the agents. An η of zero corresponds to random choices, and an η of infinity corresponds to error-free choices.

Introducing these errors affects strategy through the values. Since arm choice is determined probabilistically based on the index of each arm, a strategy is an index for each state, rather than an arm selection. For a fixed η, these can be computed numerically through backward induction, first computing the index for each state in the T^{th} period, then for each state in the $(T-1)^{st}$ period, etc. Because agents can switch from the unknown arm to the known arm and back again, there are three state variables: the number of successes observed on the unknown arm, the number of failures observed on the unknown arm, and the number of times the known arm has been selected. The index for an unknown arm is the expected value of a known arm which makes an agent indifferent between choosing the known and unknown arms in the current state.

The predictions arising from a numerical analysis of quantal response strategies are discussed in Section 4.5.6.

4.5 Experimental Design

Testing the prediction that agents with suboptimal indexes will pay more than optimal for initial information about an ambiguous arm requires two treatments: one to determine a Gittins index, and the second to determine the subject's willingness to pay for information about the value of an ambiguous arm.

	Prior (α, β)	Prior Mean	Prior Std	Periods	Gittins Index	Optimal WTP
A	(1,1)	0.50	0.29	4	0.62	0.25
B	(2.5,2.5)	0.50	0.20	4	0.56	0.21
C	(1.1,3.9)	0.22	0.17	8	0.31	0.06
D	(3.9,1.1)	0.78	0.17	5	0.83	0.04
E	(2,3)	0.40	0.20	5	0.47	0.21
F	(3,2)	0.60	0.20	8	0.69	0.24

Table 4.1: Properties of the six unknown arms used in the experiment

In each treatment, subjects will play six bandits in a random order. Each bandit will have a known, fixed horizon and no induced discounting. Each bandit will have two arms, one with an unknown (ambiguous) average payoff, and one with a known (unambiguous) average payoff.

4.5.1 Ambiguous Arms

In each treatment, subjects may choose to receive payoffs from a Bernoulli arm with an unknown (ambiguous) probability of paying off. The probability of payoff (the parameter of the Bernoulli distribution) is distributed beta(α, β), where α and β are known parameters. This probability is represented to subjects in terms of balls and urns. They are told that they are choosing between urns (arms) which contain 100 balls in some combination of red and white. When they choose an urn, one ball is drawn at random from the urn, and they earn \$1 if the ball is red and nothing if it is white. The different priors are related using tables which give the chance that there are exactly (PDF) and less than (CDF) each possible number of red balls in the unknown urn.

In separate treatments, subjects play each of six bandits with the characteristics represented in Table 4.1. These bandits were chosen to allow direct testing for three possible effects: a mean effect, a variance effect and a length effect. Bandits A and B have the same mean and same length but different variances, so their data can be compared to discover the impact of a change in variance. Bandits C and D (and E

and F) have the same variance, but different means and different numbers of periods. A mean effect can be tested by comparing data from bandits C and E (which have means below 0.5) with that from bandits D and F (which have means above 0.5). On average, these pooled data will have the same number of periods, and the same variance. A length effect can be tested by comparing the data from bandits C and F (which have length 8) and D and E (which have length 5). On average, these pooled data will have the same means and the same variance. If more than one of these effects is present, standard statistical techniques can be used to control for confounding effects.

The instructions used in each treatment are in Appendix 4.A.

4.5.2 Gittins Index Treatment

The Gittins index is elicited using the same procedure as in Chapter 3. The subject is given the table which describes the prior over the arm mean and told the number of periods in the bandit. She is then asked the minimum true mean of the known mean arm for which she would choose the known mean arm in the first period. The mean of the known mean arm is then announced, and the subject's first period arm choice is made for the subject based on her reported index: the known mean arm is chosen for her if her reported index is lower than the known mean, and the unknown mean arm is chosen otherwise. In subsequent periods, the subject can choose either the known or unknown mean arm.

The minimum mean of a known mean arm for which she would choose a known mean arm in the first period is elicited using a simple titration mechanism where the subject can respond Yes or No to a question like, "Would you choose the known mean arm this period if it paid X% of the time?" This question was repeated, with successive values of X given by a bisection algorithm, until the subject's indifference point was narrowed to the nearest percent. This value is the subject's Gittins index.

Proposition 3 applies here to prove that this mechanism is incentive compatible for the subject's true Gittins index.

4.5.3 Willingness to Pay Treatment

In the willingness to pay treatment, the subject must choose between the unknown mean arm and an arm which pays with probability one-half. Before the first period, each subject is given the opportunity to pay the experimenter to tell her the actual mean of the ambiguous arm. If she buys this information, she knows the true means of both arms, and will choose the one with the higher mean in each period. If she does not, she must choose between a known and an unknown mean arm in each period.

The amount the subject is willing to pay is elicited using a simple titration mechanism like that used to elicit the Gittins index. The subject can respond Yes or No to a question like, "Would you be willing to pay \$X.XX to learn the average of the unknown mean arm?" This question was repeated, with successive values of \$X.XX (between \$0.00 and \$1.00) given by a bisection algorithm, until the subject's indifference point was narrowed to the nearest penny. This value is the subject's willingness to pay.

Once the subject's willingness to pay is established, it is compared to a randomly determined selling price. If the subject's willingness to pay is higher than the random price, then the subject is told the true mean of the unknown mean arm and the selling price is deducted from her total payoff; if her willingness to pay is lower than the price, she is not charged, and is not told the true mean of the unknown mean arm.

Subjects' willingness to pay was assessed at the same six priors as Gittins indexes are elicited. The value of learning θ, however, also depends on λ. Theorem 12 indicates that for any fixed $\lambda < \Lambda(F, A)$, the ambiguity averse agent will pay strictly more than an ambiguity neutral agent to learn the value of θ; if $\lambda \geq \Lambda(F, A)$, then both agents would pay the same. This prediction of ambiguity aversion can be tested by holding λ fixed and changing the prior from which the value of θ is drawn. If ambiguity aversion contributes to suboptimal experimentation, agents should pay much more than optimal when high values of θ are likely, but exactly optimal when low values of θ are likely. By using $\lambda = 0.5$, the different priors will progress through

the range where ambiguity averse agents will pay more for information into that where they will pay just as much as ambiguity neutral agents.

4.5.4 Subjects

Subjects consisted of 33 Caltech undergraduates who did not necessarily have any training in economics, though many had participated in unrelated economics experiments. Experimental sessions lasted about an hour and a half, and payments averaged $18, ranging from $3 to $24.

4.5.5 Comments on Experimental Design

One possible problem with this design is that it compares the value of information based on an elicited Gittins indexes with the value of information based on a willingness-to-pay procedure. One might argue that any possible difference observed is attributable to a framing effect due to the differing procedures. There is no natural control for eliciting willingness-to-pay in the Gittins context, or vice-versa, so these effects are difficult to control for. It should be noted, however, that the phenomenon of underexperimentation is robust to procedural variances, as it appears in both the direct choice environment and the Gittins elicitation procedures. Therefore, it might be expected to be invariant to a willingness-to-pay procedure as well.

4.5.6 Predictions of Alternative Theories

In addition to ambiguity aversion, risk aversion, hyperbolic discounting and quantal response strategies make predictions about the willingness to pay treatment. However, the predictions are difficult to demonstrate analytically, so they are presented here numerically.

Risk Aversion

For two of the bandits in this experiment, risk aversion, or a distaste for the variance in the subjective payoff distribution, F, makes a different prediction than ambiguity

r	(2,3)	(1.1,3.9)
1.00	0.21872	0.06405
0.95	0.21843	0.06392
0.90	0.21815	0.06379
0.85	0.21786	0.06367
0.80	0.21758	0.06354
0.75	0.21729	0.06341
0.70	0.21700	0.06328
0.65	0.21672	0.06316
0.60	0.21643	0.06303
0.55	0.21614	0.06290
0.50	0.21585	0.06277
0.45	0.21556	0.06265
0.40	0.21527	0.06252
0.35	0.21498	0.06239
0.30	0.21469	0.06227
0.25	0.21440	0.06214
0.20	0.21411	0.06201
0.15	0.21381	0.06189
0.10	0.21352	0.06176
0.05	0.21323	0.06163

Table 4.2: Willingness to pay values for different values of r in the utility function $u(x) = x^r$ for two bandits

aversion. While ambiguity aversion predicts agents should always be willing to pay more than optimal to learn the true mean of the unknown arm, if the optimal strategy prescribes the known mean arm in the unknown arm bandit, risk aversion predicts that agents will have *lower* willingness to pay than risk neutral agents.

Table 4.2 computes the willingness to pay for agents with utility functions of the form $U(x) = x^r$, which demonstrates constant relative risk aversion. The computations presented in the table use an initial wealth level of 10, but the results are robust to all wealth levels. For the two bandits presented here,[1] more risk averse agents (with lower r) have *lower* willingess to pay than risk neutral agents.[2]

[1] Only two bandits are presented because those whose optimal strategies have the unknown arm as the initial selection are less tractable.

[2] Slight differences in willingness to pay with $r = 1$ and those in Table 4.1 arise from different precisions of algorithms in C and Excel (Table 4.1) and Mathematica (Table 4.2). These third digit differences do not affect the statistical conclusions.

The intuition for this result is that the known arm has a mean of 0.5, which maximizes the variance of the Bernoulli distribution. Therefore, the variance of the distribution over final wealth states of a strategy which selects the known arm many times is very high. For sufficiently unfavorable priors over the unknown arm, it is optimal to select the known arm in each period, and the high variance over outcomes reduces risk averse agents' utility.

The final wealth variance may be reduced by learning the value of the unknown arm. Although a risk averse agent will never select an arm with a lower mean just to reduce risk, the possibility that the true mean of the unknown mean arm is higher than 0.5, and therefore has a higher mean and lower variance than the known mean arm, induces the risk averse agent to pay to learn the mean of the unknown arm. The willingness to pay is decreasing in the level of risk aversion because more risk averse agents realize less gain from small improvements over 0.5, the most likely values of the unknown arm under the priors used here.[3]

Hyperbolic Discounting

Although hyperbolic discounting was rejected based on the data in Chapter 3, this experiment provides a second, within-experiment test. In particular, while ambiguity aversion predicts higher than optimal willingness to pay to learn the true mean of an unknown mean arm, hyperbolic discounting affects the value of both the known and unknown bandits, and the net effect is a lower than optimal willingness to pay. Table 4.3 presents the willingness to pay for each bandit for hyperbolic discounters with βs between one-half and one.

When $\beta = 1$, the willingness to pay corresponds to the optimal willingness to pay, and it decreases monotonically as β gets closer to zero.

Quantal Response Strategies

McKelvey and Palfrey's (1998) quantal response model produces a more subtle set of predictions than other models because whether the willingness to pay is predicted to

[3]This result depends on the known arm having a high-variance mean, like 0.5.

	Prior					
β	(1,1)	(2.5,2.5)	(2,3)	(3,2)	(1.1,3.9)	(3.9,1.1)
1.00	0.247	0.208	0.219	0.241	0.064	0.039
0.98	0.244	0.206	0.215	0.237	0.063	0.039
0.96	0.242	0.203	0.212	0.233	0.062	0.038
0.94	0.239	0.201	0.208	0.229	0.061	0.037
0.92	0.237	0.198	0.205	0.225	0.060	0.037
0.90	0.235	0.196	0.201	0.221	0.058	0.036
0.88	0.232	0.193	0.198	0.217	0.057	0.036
0.86	0.230	0.191	0.194	0.213	0.056	0.035
0.84	0.227	0.188	0.191	0.209	0.055	0.034
0.82	0.225	0.186	0.187	0.205	0.054	0.034
0.80	0.223	0.183	0.184	0.201	0.053	0.033
0.78	0.220	0.181	0.180	0.197	0.052	0.032
0.76	0.218	0.178	0.177	0.194	0.051	0.032
0.74	0.215	0.176	0.173	0.190	0.049	0.031
0.72	0.213	0.174	0.170	0.186	0.048	0.031
0.70	0.211	0.171	0.166	0.182	0.047	0.030
0.68	0.208	0.169	0.163	0.178	0.046	0.029
0.66	0.206	0.166	0.159	0.174	0.045	0.029
0.64	0.203	0.164	0.156	0.170	0.044	0.028
0.62	0.201	0.161	0.152	0.166	0.043	0.027
0.60	0.198	0.159	0.149	0.162	0.042	0.027
0.58	0.196	0.156	0.145	0.158	0.041	0.026
0.56	0.194	0.154	0.142	0.154	0.039	0.026
0.54	0.191	0.151	0.138	0.150	0.038	0.025
0.52	0.189	0.149	0.135	0.146	0.037	0.024
0.50	0.186	0.146	0.131	0.142	0.036	0.024

Table 4.3: Willingness to pay of a hyperbolic discounter with different βs in each bandit

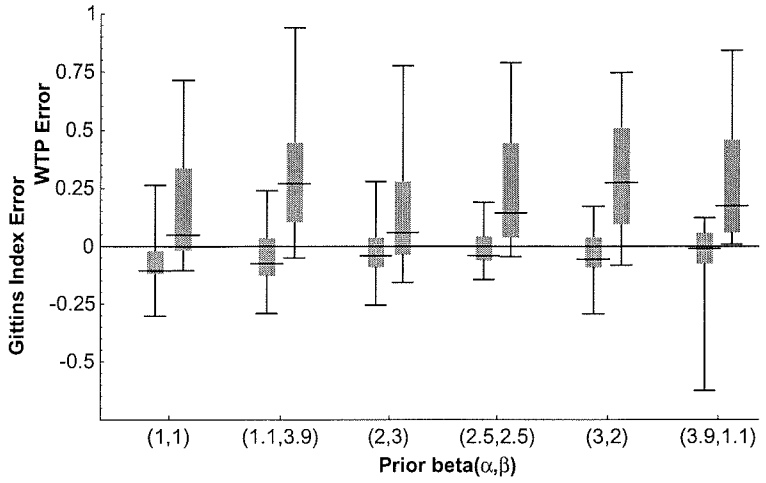

Figure 4.1: Box-and-whiskers plot of the difference between subjects' responses and the optimal Gittins index (on the left) and the optimal willingness to pay (on the right)

be higher or lower than optimal depends on the bandit and on the value of η. Table 4.4 gives the indexes for each bandit at a variety of different ηs. For each bandit, low values of η lead to lower than optimal willingness to pay, but larger values lead to higher than optimal willingness to pay, until willingness to pay returns to the optimal level at and η of infinity. The numbers in the lower section of the table represent the ηs at which the willingness to pay is maximized, and present the corresponding willingness to pay and indexes.

4.6 Results

Figure 4.1 is a box-and-whiskers plot of the data for each bandit. Gittins indexes and willingness to pay are represented as differences from optimal, with a positive difference corresponding to a higher than optimal Gittins index or willingness to pay. Each bar indicates the distribution of the data. The thin black horizontal line is at the median response, the grey box covers the middle 50%, and the long vertical lines cover 90% of the data. Note that this is the distribution of the data, and does not naturally correspond to confidence intervals of the central tendency of the data.

	Prior					
	(1,1)		(2.5,2.5)		(1.1,3.9)	
η	WTP	Index	WTP	Index	WTP	Index
0.000	0.000	0.500	0.000	0.500	0.000	0.220
0.25	0.017	0.504	0.009	0.501	0.011	0.221
0.6	0.040	0.510	0.022	0.503	0.027	0.223
1	0.066	0.516	0.037	0.505	0.045	0.226
1.25	0.082	0.520	0.046	0.506	0.057	0.227
1.75	0.112	0.528	0.063	0.508	0.078	0.230
2	0.126	0.532	0.071	0.509	0.088	0.231
3	0.176	0.547	0.103	0.513	0.122	0.237
5	0.241	0.575	0.153	0.522	0.154	0.248
7	0.270	0.597	0.187	0.530	0.151	0.258
10	0.278	0.617	0.215	0.540	0.123	0.271
15	0.270	0.630	0.229	0.553	0.083	0.289
25	0.256	0.631	0.226	0.566	0.060	0.310
50	0.246	0.623	0.215	0.568	0.061	0.322
100	0.246	0.618	0.208	0.563	0.063	0.315
500	0.246	0.616	0.208	0.560	0.064	0.308
∞	0.246	0.616	0.208	0.560	0.064	0.308
9.801	0.278	0.616				
17.282			0.230	0.558		
5.709					0.155	0.252

	(3.9,1.1)		(2,3)		(3,2)	
η	WTP	Index	WTP	Index	WTP	Index
0	0.000	0.780	0.000	0.400	0.000	0.600
0.25	0.008	0.781	0.011	0.401	0.015	0.602
0.6	0.018	0.782	0.026	0.403	0.037	0.605
1	0.029	0.784	0.043	0.405	0.060	0.608
1.25	0.035	0.785	0.053	0.407	0.074	0.610
1.75	0.046	0.787	0.074	0.409	0.101	0.614
2	0.050	0.788	0.084	0.411	0.113	0.616
3	0.062	0.791	0.121	0.416	0.157	0.623
5	0.065	0.798	0.179	0.426	0.214	0.637
7	0.054	0.804	0.217	0.436	0.239	0.649
10	0.040	0.811	0.247	0.448	0.249	0.664
15	0.032	0.820	0.264	0.464	0.245	0.683
25	0.034	0.829	0.265	0.479	0.242	0.703
50	0.038	0.831	0.247	0.482	0.244	0.706
100	0.039	0.83	0.224	0.476	0.244	0.697
500	0.039	0.826	0.219	0.472	0.241	0.691
∞	0.039	0.826	0.219	0.472	0.241	0.691
4.14	0.067	0.795				
19.367			0.266	0.473		
10.406					0.249	0.666

Table 4.4: Predicted indexes and willingness to pay of quantal response agents with different ηs, for each bandit

# Index ≤ optimal	# WTP ≥ optimal						
	0	1	2	3	4	5	6
0				1	2		
1						1	1
2							3
3						3	2
4							3
5	1					1	4
6				2	2	4	3

Table 4.5: Number of subjects who reported a higher than optimal willingness to pay in X bandits and lower than optimal Gittins indexes in Y bandits

Based on this graph, it appears the data are consistent with ambiguity aversion. In every bandit, the median Gittins index is too low and the median willingness to pay is too high. In five of the six bandits, about 90% of the willingness to pay errors are higher than the median Gittins index error.

It is also noteworthy that the level of distribution of each error remains fairly constant from one bandit to the next. There is no dramatic effect of changes in prior standard deviation (ranging from 0.169 in (1.1,3.9) and (3.9,1.1) to 0.289 in (1,1)), prior mean (ranging from (1.1,3.9) to (3.9,1.1)) or horizon length.

The rest of this section develops formal statistical tests for the patterns which appear in this graph.

Result 15 *Most subjects had higher than optimal willingness to pay and lower than optimal Gittins indexes.*

Ambiguity aversion predicts that subjects will have a willingness to pay which is too high and Gittins indexes which are too low. The frequency with which each subject made these errors is reported in Table 4.5. The columns represent the number of the six bandits in which the subject reported a higher than optimal willingness to pay, and the rows represent the number of the six bandits in which the subject reported a lower than optimal Gittins index. The number in each cell is the number of subjects who made that combination of errors. For instance, three subjects made

Prior	(1,1)	(2.5,2.5)	(1.1,3.9)	(3.9,1.1)	(2,3)	(3,2)	Pooled
Willingness to Pay							
N	33	33	33	32	33	33	197
Med overpay	0.05	0.15	0.29	0.19	0.08	0.29	0.17
# overpay	24.00	29.00	30.00	29.00	23.00	31.00	166.00
p-value	0.0045	7E-06	1E-06	2E-06	0.012	2E-07	0
Gittins Index							
N	33	33	33	33	33	33	198
Med undervalue	0.10	0.04	0.07	0.01	0.03	0.04	0.04
# too low	27.00	19.00	24.00	19.00	20.00	22.00	131.00
p-value	0.0001	0.192	0.0045	0.192	0.112	0.0278	3E-06

Table 4.6: One-tailed p-values that WTPs and Gittins indexes are optimal for each arm and for the pooled data, based on the median response

had higher than optimal willingness to pay and lower than optimal Gittins indexes in each of the six bandits in which each measurement was elicited.

An overwhelming majority of subjects fall in the lower right-hand corner of this table, where they frequently have higher than optimal willingness to pay and lower than optimal Gittins indexes. Of the 33 subjects, 25 of them have higher than optimal willingness to pay in at least five of the six bandits; 17 of the 33 have lower than optimal Gittins indexes. This pattern of response is consistent with ambiguity aversion.

Although still a majority, this rate of lower than optimal Gittins indexes is strikingly different than in Chapter 3, where nearly every subject reported a lower than optimal Gittins index in almost every first period. This could be because the scale of the indexes is much different in this experiment because the arms are Bernoulli rather than normal. The role scale effects in bandit problems may be an interesting avenue for further research.

Result 16 *The median willingness to pay is significantly higher than optimal in each bandit, and the median Gittins index is significantly lower than optimal in three of the six bandits. In the pooled data, the median willingness to pay is significantly higher than optimal and the median Gittins index is significantly lower than optimal.*

Table 4.6 presents the median overpayment and median undervaluation for each

arm, as well as for the pooled data. Because I am particularly interested in overpayment and undervaluation, these errors will both be defined as positive; a negative overpayment corresponds to underpayment, and a negative undervaluation corresponds to a higher than optimal Gittins index.

For every arm, the median overpayment is positive, meaning the median willingness to pay is higher than optimal. The third row indicates the number of overpayments which are positive. This number can be used to calculate a p-value for the hypothesis that the true median overpayment is equal to zero using Mosteller and Rourke's (1973) technique for calculating a nonparametric confidence interval for the median. Mosteller and Rourke establish a confidence interval by computing the probability that the true median is between the i^{th} and $(N-i+1)^{st}$ largest observations by computing the chance that between i and $(N-i+1)$ observations fall to the left of the median.[4] This idea can be extended to this circumstance by using the cumulative binomial to calculate the probability that, if the true median is zero, only $n \leq N$ observations are negative.

The last row presents this p-value, which is significant at all conventional levels for each arm, and extremely significant for the pooled data. Therefore, the median willingness to pay is significantly higher than optimal, consistent with the prediction of ambiguity aversion.

The second section of the table presents the same information for undervaluation. The median undervaluation is significantly positive at conventional levels for three of the six bandits. However, the median of the pooled data is highly significantly positive, consistent with the prediction of ambiguity aversion.

[4]This probability is represented by the binomial distribution and is given by

$$2\sum_{j=0}^{i}\binom{N}{j}\left(\frac{1}{2}\right)^{N}. \tag{4.61}$$

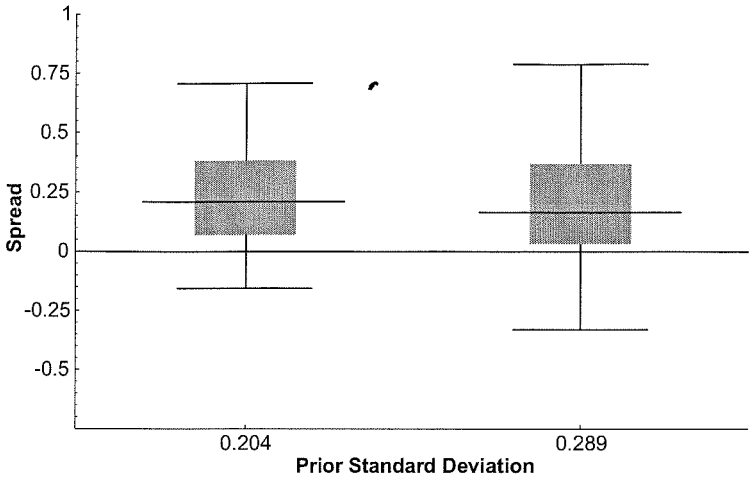

Figure 4.2: Box-and-whiskers plot of the spread for two bandits which differed only in prior variance

	(1,1)	(2.5,2.5)	p-value
Spread	0.17	0.26	1.000
Overpayment	0.05	0.15	0.317
Undervaluation	0.10	0.04	0.140

Table 4.7: Median spread, overpayment and undervaluation for the (1,1) bandit (with prior standard deviation 0.289) and the (2.5,2.5) bandit (with prior standard deviation 0.204), and p-values for the hypothesis that the two medians are the same

4.6.1 Tests of Simple Effects

The data in Table 4.6 suggest that ambiguity aversion is a significant factor in bandit problems, leading subjects to paradoxically undervalue information by having lower than optimal Gittins indexes and overvalue information by having higher than optimal willingness to pay. However, ambiguity aversion also predicts that as ambiguity increases, the undervaluation and overpayment should be more severe. However, a direct test of this prediction does not support ambiguity aversion.

Result 17 *When the mean and bandit horizon are constant, an increase in variance does not result in a significant change in subjects' Gittins indexes or willingness to pay.*

	Prior Mean $\leq .5$	Prior Mean $\geq .5$	p-value
Spread	0.20	0.26	0.336
Overpayment	0.18	0.26	0.336
Undervaluation	0.06	0.02	0.056

Table 4.8: Median spread, overpayment and underpayment for the bandits with a prior mean below one-half and those with a prior mean above one-half, and p-values for the hypothesis that the two medians are the same.

Whether variance affects the overpayment and undervaluation in the way predicted by ambiguity aversion can be tested directly by comparing the overpayment and undervaluation of the (1,1) and (2.5,2.5) bandits. Both have four periods and a prior mean of 0.5, and therefore differ only in variance. The prior standard deviation of the (1,1) bandit is 0.289 and of the (2.5,2.5) bandit is 0.204, so ambiguity aversion predicts the overpayments and undervaluations to be larger for the (1,1) bandit.

This can be tested directly by looking at the spread, the sum of the undervaluation and overpayment. Figure 4.2 is a box-and-whiskers plot of the spread for the (1,1) and (2.5,2.5) bandits. The (1,1) bandit has a slightly smaller median spread than the (2.5,2.5) bandit, the opposite of what ambiguity aversion would predict.

Table 4.7 shows the median spread for each bandit, and the p-value for the test that the medians of the two samples are the same. The continuity-corrected test statistic for the two-sample median test (Siegel and Castellan, 1988, Section 6.3) is exactly zero, leading to a p-value of one. Therefore, although the median of the higher variance bandit is lower, the difference is not significant.

The other two rows of the table, which show how overpayment and undervaluation change with ambiguity, illustrates that, although undervaluation increases with variance, overpayment seems to decrease (though not significantly).

Result 18 *When the variance and bandit horizon are constant, an increase in mean results in no change in overpayment and a borderline significant decrease in undervaluation.*

Ambiguity aversion makes a subtle prediction about how the mean will affect the spread. Although undervaluation should not be affected, the subjects' willingness to

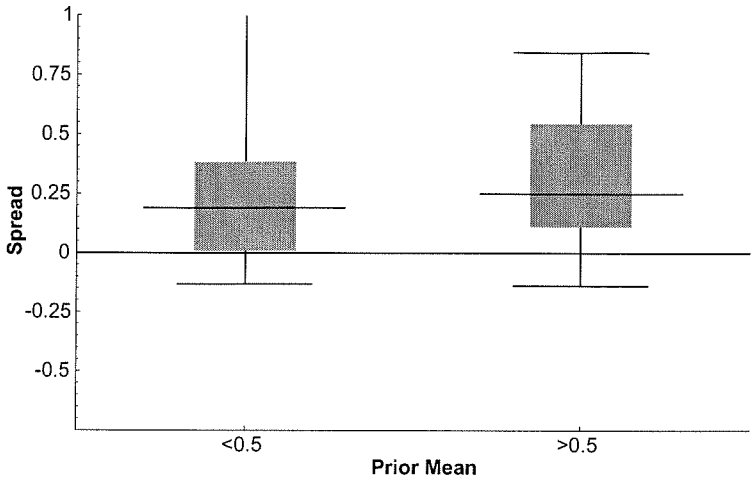

Figure 4.3: Box-and-whiskers plot of the spread for bandits which with Gittins indexes above and below the known mean arm

pay should be exactly the same as that of an ambiguity neutral agents when the mean is below 0.5. The reason for this is that, when the Gittins index is below 0.5, both the ambiguity averse and ambiguity neutral subject expects to pick the 50/50 arm in the first period (and in every period thereafter) if he does not learn the mean of the unknown arm. Therefore, there is no ambiguity in the bandit, even when there is no information. Since the two bandits with means less than 0.5 also have optimal Gittins indexes below 0.5, the ambiguity averse agents must have subjective Gittins indexes below 0.5. Therefore, when they are calculating their willingness to pay, the ambiguity averse subjects consider the two-armed bandit with the ambiguous arm to be an unambiguous problem, because they will never encounter the ambiguity in optimal play.

This suggests that the level of overpayment should increase when the mean increases from below 0.5 to above 0.5. Figure 4.3 is a box-and-whiskers plot of the spread of bandits (1.1,3.9) and (2,3) on the left and (3,2) and (3.9,1.1) on the right. There is a slight increase in the median, and in the middle two quartiles, when the mean increases.

Table 4.8 presents the median spread, overpayment and undervaluation for the bandits with a prior mean below one-half and those with a prior mean above one-half.

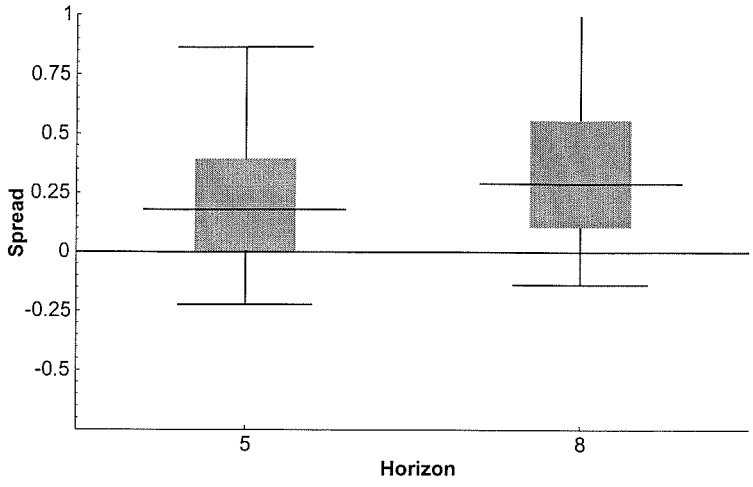

Figure 4.4: Box-and-whiskers plot of the spread for bandits with different horizons

Although the median spread for the means above 0.5 is higher than that for the means below 0.5, the difference is not significant; nor is the difference in median overpayment. Interestingly, undervaluation is slightly lower than for priors with higher means, and this difference is borderline significant.

It is possible that this lack of increase in the spread is attributable to ambiguity aversion, if subjects' Gittins indexes for the arms with means above 0.5 are, due to ambiguity aversion, also below 0.5. The reported Gittins indexes, though not infrequently below the prior mean of the ambiguous arm, are rarely that much lower. Therefore, the lack of difference is probably not attributable to ambiguity aversion.

The key to the lack of increase may lie in the the fact that the premise of the prediction does not hold: agents do not have optimal willingness to pay in the arms with means below 0.5. Table 4.6 shows that bandits (1.1,3.9) and (2,3) (and (1,1) and (2.5,2.5)) have significantly higher than optimal willingness to pay. Considering why this is so may provide insight into why there is no mean effect.

Result 19 *When the variance and mean are constant, an increase in horizon results in an increase in both overpayment and undervaluation.*

Ambiguity aversion does not hold any role for the horizon, so any change in spread is not attributable to ambiguity aversion but may provide some insight into subjects'

	5 Periods	8 Periods	p-value
Spread	0.18	0.30	0.011
Overpayment	0.14	0.25	0.067
Undervaluation	0.02	0.07	0.056

Table 4.9: Median spread, overpayment and underpayment for the bandits with a five period horizon and those with an eight period horizon, and p-values for the hypothesis that the two medians are the same

decision process.

Figure 4.4 is a box-and-whiskers plot of the spread of the (3.9,1.1) and (2,3) bandits on the left and the (1.1,3.9) and (3,2) bandits on the right. There is a slight increase in the median, and in the middle two quartiles, with the longer horizon.

Table 4.9 presents the median spread, overpayment and undervaluation for the bandits with a five period horizon and those with an eight period horizon. There is a statistically significant increase in spread from the five period to the eight period horizon, generated by borderline significant increases in both overpayment and undervaluation.

Both the optimal Gittins index and the optimal willingness to pay are increasing in the horizon. These results suggest the subjects are too sensitive to the increase in willingness to pay, but not sensitive enough to the increase in the Gittins index.

This sort of relationship might arise from a simplification of Equation 4.55.[5] In this experiment γ_1 is simply the number of periods in the bandit. Subjects may try to approximate $V()$ by some linear function of the number of periods, rather than solve the dynamic programming problem. If this approximation yielded an average per-period value which was lower than the true per-period average value (which would be consistent with observing lower than optimal Gittins indexes), subjects would have an average per-period willingness to pay which was higher than optimal. Furthermore, their simple model would be linearly increasing in the horizon, so the amount of the overpayment would be increasing in the horizon.

[5] After one session, a subject described to me exactly Equation 4.55 (these are Caltech students) and then asked, "but how do you compute the value of not having the information?"

	Constant	Prior Mean	Prior Std	Log(Horizon)
Spread	-0.03	0.12	-0.28	0.19
	(-0.14)	(0.75)	(-0.54)	(2.59)
Overpayment	0.07	0.07	-0.52	0.16
	(0.31)	(0.64)	(-1.31)	(2.08)
Undervaluation	0.12	0.09	0.28	0.03
	(-0.85)	(0.57)	(0.84)	(1.25)

Table 4.10: Results of regressions of spread, overpayment and undervaluation on prior mean and variance and the log of horizon (t-statistics are in parentheses)

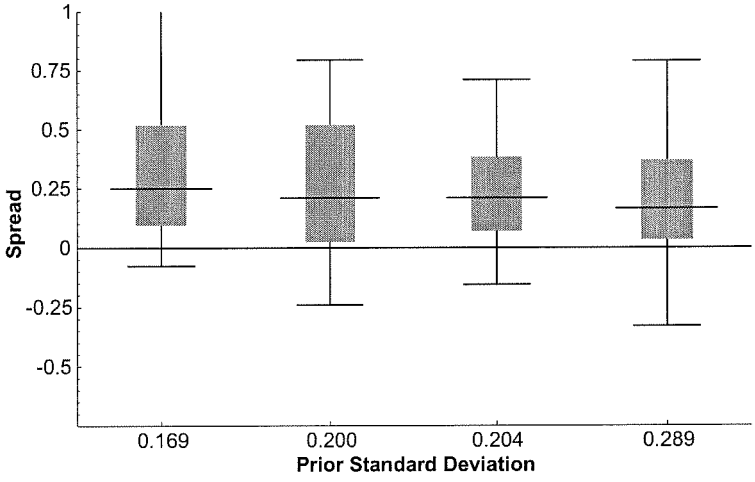

Figure 4.5: Box-and-whiskers plot of the spread for all bandits ordered by prior standard deviation

4.6.2 Testing for Multiple Effects

The previous section used the design of the experiment to test simple hypotheses about the impact of changes in prior mean and variance and the horizon. However, this approach throws out some of the data, as not all six bandits are used in any of these tests. This section uses a simple linear model to test for these effects in the whole sample.

Result 20 *When tested on the entire sample, only the horizon effect is significant.*

The simple test for a change in the spread related to a change in the variance only used one-third of the sample, and only two variances. Figure 4.5 shows the

distribution of the data in the bandits with every variance used. As with the simple test, there is no dramatic trend toward larger spreads with higher standard deviations. However, this larger sample does not carefully control for mean and horizon effects.

The regressions in Table 4.10 use the whole sample and control for linear mean and horizon effects, and come to the same conclusion: the spread does not vary significantly with changes in the variance. This is an important prediction of ambiguity aversion, and without support for it, it is difficult to say conclusively that the data in Table 4.6 are actually attributable to it. However, it may be that the small range of standard deviations used here is not large enough to generate a detectable difference.

The mean effect is also still not significant, but the horizon effect remains significant. Given that the horizon effect is consistent with a particular type of systematic error in the computation of the value function and the Gittins index, it seems possible that overpayment is attributable to the same cause as undervaluation. If subjects do not sufficiently compensate for solving a shortened version of the problem, undervaluation, and thus overpayment, are explained by horizon truncation, the model which was consistent with the data presented in Chapter 3.

4.6.3 Results of Alternative Theories

Risk Aversion

Risk aversion and ambiguity aversion predict opposite deviations from optimality in the willingness to pay treatment for priors (2,3) and (1.1,3.9).

Result 21 *With priors (2,3) and (1.1,3.9), subjects' willingness to pay is significantly higher than optimal, inconsistent with risk aversion.*

Table 4.6 tests whether the median willingness to pay for each bandit is different from the optimal value. For both bandits in which ambiguity aversion and risk aversion make different predictions, subjects' willingness to pay is significantly higher than optimal. Since risk aversion predicts willingness to pay will be lower than optimal, the data are not consistent with risk aversion being the primary factor leading to higher than optimal willingess to pay.

Hyperbolic Discounting

Table 4.3 shows that hyperbolic discounting predicts that agents' willingness to pay will be lower than optimal.

Result 22 *Subjects' willingness to pay is significantly higher than optimal, inconsistent with hyperbolic discounting.*

Table 4.6 tests whether the median willingness to pay for each bandit is different from the optimal value. For each bandit, willingness to pay is significantly higher than optimal. This is not consistent with hyperbolic discounting. Therefore, hyperbolic discounting is not the primary cause of higher than optimal willingness to pay.

Quantal Response

Quantal response makes different predictions about whether indexes and willingness to pay will be higher or lower than optimal depending on the prior and on the value of η. Rather than compute a particular value of η for all bandits, quantal response is tested here by comparing observed willingness to pay and Gittins indexes to those predicted at the η which predicts the maximum possible willingness to pay for each bandit.

Result 23 *Subjects' willingness to pay is significantly higher than the highest possible willingness to pay predicted by quantal response.*

Table 4.11 tests whether the median observed willingness to pay is significantly larger than the maximum possible willingness to pay under quantal response.

The upper section of the table presents the median amount more than the maximum willingness to pay possible under quantal response in the data, the number of agents who were willing to pay more than this maximum, and the p-value of the one-tailed test that the median of the sample is greater than the maximum quantal response willingness to pay. In four of the six bandits, as well as in the pooled data, this difference is highly statistically significant.

Prior	(1,1)	(2.5,2.5)	(1.1,3.9)	(3.9,1.1)	(2,3)	(3,2)	Pooled
Willingness to Pay							
N	33	33	33	32	33	33	197
Med overpay	0.02	0.13	0.20	0.16	0.03	0.28	0.14
# overpay	20.00	29.00	26.00	28.00	17.00	31.00	151.00
p-value	0.112	6E-06	4E-04	1E-05	0.431	2E-07	4E-14
Gittins Index							
N	33	33	33	33	33	33	198
Med undervalue	0.10	0.04	0.01	-0.02	0.03	0.02	0.02
# too low	27.00	19.00	20.00	10.00	20.00	19.00	115.00
p-value	1E-04	0.192	0.112	0.988	0.112	0.192	0.011

Table 4.11: One-tailed p-values that median willingness to pay and Gittins indexes are generated by quantal response agents with willingness to pay-maximizing ηs, for each arm and for the pooled data

The lower section of the table reports the median amount less than the Gittins index with the η generating the maximum willingness to pay the data were, as well as the number which were too low, and the p-value of a one-tailed test that the population median index was equal to that predicted by quantal response. This hypothesis is rejected in two of the six bandits at conventional levels of significance.

The willingness to pay is too high to be explained by any value of η in four of the six bandits, and in a fifth, the index corresponding to a willingness to pay which cannot be rejected is strongly rejected. Therefore, behavior in five of the six bandits is inconsistent with quantal response. Furthermore, this test places no restrictions on the model, allowing a different η for each bandit. These values vary significantly, and any single η will predict higher than optimal willingness to pay in some bandits, but lower than optimal willingness to pay in others. Therefore, patterns of willingness to pay choice in the data cannot be explained by quantal response.

4.7 Discussion

This chapter showed how a distaste for ambiguity, the variance of the prior distribution of the mean, might lead to the lower than optimal Gittins indexes implied

by results in search problems, as well as in the Chapters 2 and 3. Ambiguity averse agents dislike receiving payoffs from ambiguous process, and therefore are willing to accept less valuable unambiguous alternatives than would be an ambiguity neutral agent. This means their Gittins indexes are lower than optimal.

Lower than optimal Gittins indexes make it seem as though ambiguity averse agents do not place enough value on the information they gain from experimentation. It is paradoxical, then, that if the problem is reframed to explore how much they would be willing to pay for information about an ambiguous alternative, agents are willing to pay more than an ambiguity neutral agent, appearing to overvalue information.

This surprising prediction of ambiguity aversion is unique among theories which assume that agents correctly formulate and solve the dynamic programming problem. Other theories of systematic deviations, such as hyperbolic discounting and risk aversion, affect the calculation of the value function in both the ambiguous and unambiguous cases, so the difference in willingness to pay will not be generated. Of course, if agents do not correctly formulate and solve dynamic programming problems, almost any effect can be predicted.

The experiment presented here tested this surprising prediction, using one treatment which elicited subjects' Gittins index on six bandits, and another treatment which elicited subjects' willingness to pay on the same six bandits. The basic results are exactly as ambiguity aversion predicts: subjects have significantly lower than optimal Gittins indexes, and are willing to pay significantly more than optimal to learn the true mean of the ambiguous arm.

Importantly, other theories which might explain the initial underexperimentation and lower-than-optimal Gittins indexes make different predictions than ambiguity aversion in the willingness to pay treatment for some or all of the bandits. That behavior in each bandit is consistent with ambiguity aversion suggests it is also inconsistent with other candidate explanations. Hyperbolic discounting is also rejected here as its prediction that willingness to pay will be smaller than optimal is not realized. Evidence also contradicts risk aversion's prediction of smaller than optimal willingness to pay on bandits with optimal strategies of selecting the known arm ini-

tially. Finally, the deviations from optimality are larger than could be predicted by quantal response, even allowing a different η for each bandit.

However, not every prediction of ambiguity aversion receives strong support. It also predicts that undervaluation and overpayment should increase with the ambiguity of the alternative. Similarly, there should be an increase in overpayment as the Gittins index moves from below 0.5 to above 0.5. However, neither of these predicted effects is significant. Instead, the only significant change in overpayment and undervaluation arises from increases in the number of periods in the bandit.

The increase in the spread as the horizon increases is mostly attributable to an increase in overpayment. Why would agents be willing to pay more to remove ambiguity from a longer bandit? They may be willing to pay more to avoid having to solve the longer problem explicitly, giving up some money to save mental computation cost. However, this does not explain why Gittins indexes are lower than optimal.

Rather, because willingness to pay is higher than optimal when means are below 0.5, it appears likely that at least some of the suboptimality observed is attributable to a difficulty in solving the dynamic programming problem. If horizon truncation results in a low value for the ambiguous bandit, an intuitive simplification of Equation 4.55 will result in overpayment increasing with the horizon, without there necessarily being an impact on undervaluation.

That difficulty solving the dynamic programming problems is causing undervaluation and overpayment can be tested in a couple ways. A simple to solve bandit, such as one in which the unknown arm either always paid or never paid (with known probabilities) could test subject's intuitive dynamic programming ability. In retrospect, it would have been nice to include such a bandit among those used in this experiment. Alternately, subjects could be brought in for a thorough training session on dynamic programming which could elaborate on the direction and magnitude of the effects discussed in the strategy section of the instructions.

Finally, this chapter accomplished a broader goal of demonstrating how simple behavioral models can be integrated into larger bodies of theory to produce testable predictions based on sound psychology. Although there are a number of different

models of ambiguity aversion, the Kahn and Sarin model leverages the second order probability available to bandit agents. With some interpretation, the behavioral model was integrated into the sophisticated mathematics of bandits to derive the surprising and testable prediction that agents will have lower than optimal Gittins indexes, but be willing to pay more than optimal to eliminate ambiguity. Hopefully the exercise of developing the theory and testing it will inspire others to integrate formal behavioral models into their theories.

4.A Instructions

4.A.1 Gittins Index Treatment

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.

How You Make Money

You will earn one dollar each time a red ball is drawn from an urn you choose. In each period, you may choose one of two urns: one with 100 balls in some known mixture of red and white, and one with 100 balls in some unknown mixture of red and white. When you choose an urn, one ball will be randomly drawn from that urn and it determines your payoff: one dollar if it is red, nothing if it is white.

You will be paid in cash for all of your earnings in excess of $28.

Order of the Experiment

This experiment will proceed as a number of rounds. Each round will have several periods. The number of periods in each round will be announced at the beginning of that round.

At the beginning of each round, the computer will randomly determine the number of red balls in the known mixture urn and in the unknown mixture urn. Before you learn the number of red balls in the known mixture urn, you will be asked the smallest number of red balls which would lead you to choose the known mixture urn in the first period. Your first period choice will be made for you by the computer based on your response. The computer will choose the known mixture urn for you if it contains more red balls than your smallest number, and the unknown mixture urn for you if the known mixture urn contains fewer red balls than your smallest number.

After the first period, you may select either urn. In each period, you will have to trade off selecting the known mixture urn with learning more about the number of red balls in the unknown mixture urn.

Urns

At the beginning of each round, you will be told the number of red balls in the known mixture urn. It may contain any number between zero and 100 red balls.

You will not know the mixture of red and white balls in the unknown urn. It may contain any number between zero and 100 red balls. However, not all mixtures are equally likely. The likelihood of different proportions of red balls is represented on the urn tables.

The number of red balls in both urns will remain constant from one period to the next, but will change at the beginning of each round.

Urn Tables

The chance that the unknown mixture urn contains a given number of red balls is represented in tables like the Practice Urn Table you have been given. A new table will be distributed at the beginning of each round.

The first column of the table shows a possible number of red balls. The second column indicates the chance that there are exactly X number of red balls in the unknown mixture urn. This is also illustrated in the graph below the table with a

solid black line (read on the left axis). For instance, there is a 1% chance there are exactly 25 red balls (and 75 white balls) in the unknown mixture urn.

The third column of the table shows the chance that there are **at least** X red balls in the unknown mixture urn. This is illustrated in the graph below the table with a dashed lack line (read on the right axis). For instance, there is a 49.08% chance that there are fewer than 50 red balls in the unknown mixture urn.

Known Mixture Urn Cutoff

In the first period, the computer will ask you "Would you choose the known mixture urn in this period if it contained [Number] red balls?" 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 known mixture urn has been narrowed down to the nearest ball.

You should answer these questions carefully because your first period urn choice will be made for you based on your answers. The computer assumes you will choose the known mixture urn for all numbers of red balls larger than the cutoff, and the unknown mixture urn otherwise. Therefore, it will automatically choose the known mixture urn if it has more red balls than your cutoff, and the unknown mixture urn if the known mixture urn has fewer red balls than your cutoff.

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.

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 minimum value for which you would choose the known mixture urn; recent periods will be added to the top of the list, though earlier periods will still be accessible by scrolling down.

The Information Panel

The top of the three panels on the right side provides you with information on the current period, the total number of periods and your total payoff. It also provides a "Best Guess" at the number of red balls in the unknown mixture urn. In the first period, it shows the average number of red balls that would be an urn based on the urn table. Once the unknown mixture urn is chosen, the "Best Guess" uses a law of probability called Bayes' rule. Bayes' rule uses the chance the observed combination of red and white balls arose from each possible mixture, and the chance of each mixture from the urn table, to determine the most likely average number of red balls in the unknown mixture urn, given the available information.

The Urn Choice Panel

Please look at the middle of the three right-hand 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.

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.

Summary

1. At the beginning of the round, the Experimenter will distribute an urn table and announce the number of periods in the round.

2. The computer will randomly select the number of red balls in the known mixture urn, and in the unknown mixture urn.

3. You will be asked a series of questions to determine your known mixture urn cutoff, the minimum number of red balls in the known mixture urn for which

you would choose it that period.

4. The computer will automatically choose the known mixture urn for you if its actual number of red balls is higher than your cutoff, and the unknown mixture urn otherwise.

5. The computer will randomly draw a ball from your chosen urn and announce your payoff: one dollar if the ball is red and nothing if it is white.

6. Fill in the record section of the urn table.

7. Wait for the experimenter to announce the beginning of the next period.

8. Choose between the known and unknown mixture urns, and return to Step 5.

Strategy

You have a chance to receive a payoff when you select either urn, but when you select the unknown mixture urn, you also gain some information about the number of red balls it contains, which may help you in future periods. At any point, your best guess may be higher or lower than the actual number of red balls in the urn. By trying the unknown mixture urn, you may learn it has more red balls than the known mixture urn, information you can use to improve your chance of getting a red ball in future periods; if the unknown mixture urn does not have more red balls, you can choose the known mixture urn in future periods.

This possibility of learning the unknown mixture urn is better than your initial best guess means it is sometimes advantageous to select the unknown mixture urn even when the known mixture urn contains more red balls than your best guess at the number in the unknown mixture urn. Whether it is worth experimenting with the unknown mixture urn depends on the difference between your best guess and the number of red balls in the known mixture urn, the number and color of the balls you have observed from the unknown mixture urn, the chance that the actual number of red balls in the unknown mixture urn is each value higher than your best guess

(based on the urn table) and number of periods you have left to benefit from learning the unknown mixture urn has more red balls than your initial best guess.

4.A.2 Willingness to Pay Treatment

You are about to participate in an experiment designed to provide insight into certain features of 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.

How You Make Money

You will earn one dollar each time a red ball is drawn from an urn you choose. In each period, you may choose one of two urns: one with 50 red balls and 50 white balls, and one with 100 balls in some unknown mixture of red and white. When you choose an urn, one ball will be randomly drawn from that urn and it determines your payoff: one dollar if it is red, nothing if it is white.

You will be paid in cash for all of your earnings in excess of $28.

Order of the Experiment

This experiment will proceed as a number of rounds. Each round will have several periods. The number of periods in each round will be announced at the beginning of that round.

At the beginning of each round, the computer will randomly determine the number of red balls in the unknown mixture urn. You will then be asked how much you would be willing to pay to learn the number of red balls in the unknown mixture urn. If you are willing to pay more than the computer's randomly determined selling price, the computer will tell you then number of red balls in the unknown mixture urn and deduct the selling price from your total payoff.

If you are not willing to pay more than the computer's randomly determined selling price, you will not be charged, but you will only be able to learn about the

number of red balls in the unknown mixture urn by choosing it. In each period, you will have to trade off choosing the 50-50 urn with learning more about the number of red balls in the unknown mixture urn.

Urns

In each period you will be choosing between two urns. One urn will always be a 50-50 urn, containing 50 red balls and 50 white balls. On average, this urn will pay one dollar half the time.

You will not know the mixture of red and white balls in the unknown urn. It may contain any number between zero and 100 red balls. However, not all mixtures are equally likely. The likelihood of different proportions of red balls is represented on the urn tables.

The number of red balls in the unknown mixture urn will remain constant from one period to the next, but will change at the beginning of each round.

Urn Tables

The chance that the unknown mixture urn contains a given number of red balls is represented in tables like the Practice Urn Table you have been given. A new table will be distributed at the beginning of each round.

The first column of the table shows a possible number of red balls. The second column indicates the chance that there are exactly X number of red balls in the unknown mixture urn. This is also illustrated in the graph below the table with a solid black line (read on the left axis). For instance, there is a 1% chance there are exactly 25 red balls (and 75 white balls) in the unknown mixture urn.

The third column of the table shows the chance that there are **at least** X red balls in the unknown mixture urn. This is illustrated in the graph below the table with a dashed lack line (read on the right axis). For instance, there is a 49.08% chance that there are fewer than 50 red balls in the unknown mixture urn.

Willingness to Pay (WTP) for Information

At the beginning of each round, you will have the opportunity to buy from the experimenter the number of red balls in the unknown mixture urn. You may want to pay for this information because it tells you which urn has more red balls, and therefore is more likely to pay one dollar. If you do not have this information, you can learn whether there are many red balls in the unknown mixture urn only by choosing it and observing your payoffs. On the other hand, you do not want to pay more for this information than you can gain by having it.

Using the Computer to Purchase Information

After you have been shown the urn table and learned the number of periods in the round, you will be asked "Would you be willing to pay $X.XX to learn the number of red balls in the unknown mixture urn?" If you would be willing to pay that amount, click "Yes," if not, click "No." The computer will ask a series of these questions, with different values, until it has narrowed the amount you are willing to pay to the nearest cent.

Once it has determined how much you are willing to pay, the computer will compare your value to the randomly determined price at which it will sell the information. If your WTP is higher than the computer's price, the computer will tell you the number of red balls in the unknown mixture urn and deduct its price (not your WTP) from your total payoff. If your WTP is lower than the computer's price, you will not be charged, but you will not be told the number of red balls in the unknown mixture urn.

Be careful in selecting your WTP. If you enter a value which is higher than you are really willing to pay, you may have to pay more for the information than you want; if you enter a value which is lower than you are really willing to pay, you may not receive the information when the computer would be willing to tell you for a price you would be willing to pay.

Using the Computer to Choose an Urn

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.

The History Panel

The long vertical panel on the left will contain your playing history. For each period, it will show your choice and the payoff you received; recent periods will be added to the top of the list, though later periods will still be accessible by scrolling down.

The Information Panel

The top of the three panels on the right side provides you with information on the current period, the total number of periods and your total payoff. It also provides a "Best Guess" at the number of red balls in the unknown mixture urn. In the first period, it shows the average number of red balls that would be in an urn based on the urn table. Once the unknown mixture urn is chosen, the "Best Guess" uses a law of probability called Bayes' rule. Bayes' rule uses the chance the observed combination of red and white balls arose from each possible mixture, and the chance of each mixture from the urn table, to determine the most likely average number of red balls in the unknown mixture urn, given the available information.

The Urn Choice Panel

The middle of the the three right-hand panels is where you indicate your choice of urn in 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 generate a Random Value and calculate your payoff for the period.

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.

Summary

1. At the beginning of the round, the Experimenter will distribute an urn table and announce the number of periods in the round.

2. The computer will randomly select the number of red balls in the unknown mixture urn using the urn table.

3. The computer will ask you a series of questions to determine the maximum price you are willing to pay to learn the number of red balls in the unknown mixture urn.

4. The computer will compare your WTP to a random selling price.

   (a) If your WTP is higher than the selling price, the computer will tell you the number of red balls in the unknown mixture urn and deduct the selling price (not your WTP) from your total payoff.

   (b) If your WTP is lower than the selling price, the computer will not tell you the number of red balls in the unknown mixture urn.

5. The experimenter will instruct you to choose an urn.

6. The computer will randomly draw a ball from your chosen urn and announce your payoff: one dollar if the ball is red and nothing if it is white.

7. Fill in the record section of the urn table.

8. Wait for the experimenter to announce the beginning of the next period.

9. Choose between the known and unknown mixture urns, and return to Step 6.

Strategy

You have a chance to receive a payoff when you select either urn, but when you select the unknown mixture urn, you also gain some information about the number of red balls it contains, which may help you in future periods. At any point, your best guess

may be higher or lower than the actual number of red balls in the urn. By trying the unknown mixture urn, you may learn it has more than 50 red balls, information you can use to improve your chance of getting a red ball in future periods; if the unknown mixture urn has fewer than 50 red balls, you can choose the 50-50 urn in future periods.

This possibility of learning the unknown mixture urn is better than your initial best guess means it is sometimes advantageous to select the unknown mixture urn even when your best guess at the number in the unknown mixture urn is less than 50. Whether it is worth experimenting with the unknown mixture urn depends on the difference between your best guess and 50, the number and color of the balls you have observed from the unknown mixture urn, the chance that the actual number of red balls in the unknown mixture urn is each value higher than your best guess (based on the urn table) and number of periods you have left to benefit from learning the unknown mixture urn has more red balls than your initial best guess.

Feel free to earn as much money as you can. Are there any questions?

Chapter 5 Implications and Conclusions

The first objective of this thesis was to determine if the suboptimality which has been observed in search problems extends to the more general set of bandit problems.

5.1 Undervaluation of Information

Undersearch manifests itself both as failure to search when an optimally searching agent would and as lower than optimal reservation wages. Analogously, in bandit problems, undervaluation of information manifests itself both as a failure to experiment when an optimally experimenting agent would and as lower than optimal Gittins indexes. Both of these phenomena occur in laboratory bandits.

The choice data presented in Chapter 2 show that experimental subjects quickly extinguish arms which do not perform well initially, while an optimally experimenting agent would continue experimentation to ensure that the bad initial observations were not simply bad draws from a good payoff distribution. The consequence of this underexperimentation is that agents often play the highest average payoff arm with lower frequency than optimally experimenting agents; they lose money in expectation.

The loss arises from the cases where high average payoff arms yield bad initial payoffs and actual subjects abandon them when optimally experimenting agents would continue to try them and might, through additional experimentation, learn that they are better and worth choosing in the future. It is the nature of bandit problems, however, that this loss is only in expectation; agents will not not always be worse off, and sometimes even better off, for underexperimentation.

The Gittins index data presented in Chapters 3 and 4 corroborates the choice data in the claim that subjects do not value information optimally. As a generalized reservation wage, a subject's Gittins index represents a combination of the expected payoff from an arm and the present discounted value of the increase in expected

future payoffs arising from information gained by experimenting with that arm in the current period. In both experiments which directly elicited subjects' Gittins indexes, first period Gittins indexes were significantly below optimal.

When arms have normal payoff distributions and normal priors, as in Chapter 3, nearly all choices of first period Gittins index are below the optimal value. Fewer first period Gittins indexes are below optimal when the arms give Bernoulli payoffs and have beta priors, as in Chapter 4. However, in every bandit, the median index was below optimal, in most cases highly significantly.

Therefore, looking at choice data and at Gittins indexes directly, these three distinct experiments suggest that experimental subjects value the information they obtain from experimentation less than the theory suggests they should. This extends the undersearch result to the more general bandit environment.

5.2 Regularities Emerging from Experimental Data

Having established that experimental subjects do not behave optimally in bandits, the next natural question is how are they behaving? A careful understanding of systematic deviations from optimality can help economists, companies and the government fashion strategies and policies to best respond to people's actual experimentation strategies.

Section 5.1 presented data supporting the claim that agents undervalue the information they gain from experimentation in the first period of a bandit. However, the complete characterization of the data is more subtle. Rather than just undervaluing information, subjects consistently overvalue additional information once they have done some initial experimentation.

In the choice data in Chapter 2, although subjects too quickly dismiss arms yielding one or two bad initial payoffs, they continue to switch among the remaining arms at a rate which decreases only slowly. The optimal rate of switching decreases quickly, eventually coninciding with the subjects' rate of experiementation. However, the optimal rate continues to decrease faster than the subjects'; the subjects experiment

too much in later periods. This can be seen in Figure 2.6, where agents are mixing among more arms than would an optimal agent.

This later overexperimentation may be an attempt to balance the competing interests of gathering information and exploiting the arm with the highest expected value. Rather than following an index strategy, subject may devote some periods to experimentation, and some to exploitation. However, they would be better off if they experimented early, because then they would be able to act on the information they acquired in more periods.

However, this overvaluation of additional information is more fundamental than a simple rule of thumb for experimenting in the choice problem. It also appears in Gittins index data. In Chapter 3, Gittins indexes are elicited for several periods in each bandit. Although information value ratios are initially significantly below one, they increase significantly in later periods. This suggests that increasing experimentation, relative to optimal, is a systematic feature of how people approach bandit problems in the laboratory.

Exactly how this combination of too-low initial indexes and too-high later indexes plays out in naturally occurring bandits is unclear. If choices are determined only by indexes, as they were in Chapter 3, then suboptimal initial indexes may prevent agents from ever making the initial choice of an arm. Without this initial choice, they may never reach the point where their indexes are higher than optimal.

One outcome is suggested by the choice data: subjects may simply not experiment enough initially, but then experiment too much later. This is not payoff maximizing, but may be better than playing an index strategy with suboptimal index values; at the very least, it improves the chance of converging to the arm with the highest average payoff.

However, overexperimentation in the choice data may be caused by effects not necessarily present in naturally occurring bandits. A "white coat" effect, or even just boredom, could lead subjects to overexperiment in the choice environment. However, these effects probably do not act in a Gittins elicitation environment, and are unlikely to affect choices outside the laboratory. Whether or not agents in naturally occurring

bandits conduct the initial experiments is an empirical question, and an important avenue for future research.

5.3 Explanations for Suboptimal Bandit Behavior

Two explanations given in the search literature to explain undersearch may also lead to underexperimentation in the bandit environment. Risk aversion and unobservable experimentation cost would lead subjects to experiment less than risk-neutral optimality and to have lower than risk-neutral optimal Gittins indexes.

The experiment in Chapter 2 controlled for risk aversion by having equally risky arms and for experimentation cost by not imposing any. These explanations are clearly rejected by this data, as subjects do not initially experiment enough despite the absence of experimentation cost and risk-based differences in the arms.

This result is robust to the Gittins elicitation environment. The experiment in Chapter 3 also had constant risk arms and imposed no experimentation cost. Even with these factors controlled, initial Gittins indexes are lower than optimal.

5.3.1 Psychologically-based Models of Suboptimal Bandit Behavior

In addition to risk aversion and unobservable experimentation cost, three models based on stylized facts about human behavior were considered. The motivation for these models is that some aspect of human psychology may introduce a systematic bias into the way agents solve bandit problems. Identifying a systematic bias, and understanding that it is attributable to a particular psychological cause, can serve as the basis for policies, strategies and institutions which help agents improve their welfare.

Hyperbolic Discounting

Hyperbolic discounting attributes suboptimal Gittins indexes to the discount sequence. Because hyperbolic discounters place relatively more weight on the current period than geometric discounters, the relative value of the future payoffs which benefit from present experimentation is smaller. Therefore, hyperbolic discounters have smaller than optimal Gittins indexes, and will tend to select the arm with the highest expected value rather than experiment with arms with lower expected values, but about which less is known. However, because the hyperbolic discounters are not time consistent, they will regret not experimenting, and thus be worse off.

The data from Chapter 3 demonstrate that information value ratios are increasing from below one to above one as information is gained. Since the information value ratio is always on the same side of one as the hyperbolic discount factor β, this implies that a characteristic of the agent is changing as the game is played. Hyperbolic discounting does not allow for this. Therefore, hyperbolic discounting is rejected in laboratory bandits. However, since there is no meaningful time structure in this experiment, it does not rule out the possibility that hyperbolic discounting plays a role in economically important naturally occurring bandits, or even other laboratory experiments, with significant amounts of time between choices.

Ambiguity Aversion

Ambiguity aversion holds that variance in the second order probability distribution, the prior over the means of the payoff distribution, leads to suboptimal Gittins indexes. Because agents dislike second order variance, they are willing to accept a lower-valued "unambiguous equivalent" arm than would be an ambiguity neutral agent.

Chapter 4 proves that ambiguity averse agents will have lower than optimal Gittins indexes, but will also be willing to pay more than optimal to be told the true mean of an ambiguous arm. This asymmetry arises because ambiguity averse agents value the unambiguous bandit as an ambiguity neutral agent would, so the difference between

the value functions for the unambiguous bandit and the ambiguous bandit is larger when the value of the ambiguous bandit is affected by ambiguity aversion.

The data in Chapter 4 are consistent with this model. Agents have lower than optimal Gittins indexes, and are willing to pay more than an ambiguity neutral agent. However, small variations in the mean and in the variance of the second order probability distribution do not result in predicted significant changes in the level of the Gittins index or the amount of willingness to pay.

Features of the data from other experiments are also consistent with ambiguity aversion. Subjects' initial hesitancy to experiment in the choice data can be explained by ambiguity aversion, as can the lower than optimal initial Gittins indexes in Chapter 3. Furthermore, as information is gained about uncertain alternatives, the variance of the Bayesian posterior of the second order distribution decreases, so the Gittins index of an ambiguity averse agent should converge to that of an ambiguity neutral agent. Therefore, ambiguity aversion predicts the observed increase in the elicited information value ratios.

Horizon Truncation

Horizon truncation is a third behavioral explanation which appeals to an intuitive notion of how agents might solve dynamic programming problems (rather than to a well-documented psychological phenomenon like hyperbolic discounting or ambiguity aversion). It holds that agents approximate the solution to a dynamic programming problem by solving a short horizon version of the problem explicitly, and then adding an "adjustment factor" to compensate for the periods they omitted in the truncated horizon version of the problem.

That there is not enough initial experimentation in the choice data, and that initial Gittins indexes are suboptimal, is consistent with horizon truncation if the adjustment factor is too small; that there is too much later experimentation, and that later Gittins indexes are supraoptimal, is consistent with horizon truncation if the adjustment factor is larger than optimal. Therefore, the overall paths of the data can be explained by a version of horizon truncation which has an adjustment

factor which is initially too small, but does not decrease quickly enough in the face of additional information. Furthermore, the information values are decreasing, which suggests they satisfy an intuitive restriction on the adjustment factors.

5.3.2 Sampling Patterns

Although the aggregate data are consistent with both ambiguity aversion and horizon truncation, it is important to remember that individual period-by-period choices are not consistent with either of these theories, or any of the others considered here. In the choice environment, subjects experiment with simple cycling algorithms which are sensitive to only extreme payoffs. These algorithms do not produce choice patterns which look like those from an index strategy. In aggregating the data, individual cycling patterns can be averaged out to gain a sense of how much value the population attributes to the information gained from experimenting with each arm.

Therefore, these models can be thought of as modeling either the "average" strategies of a heterogeneous population, or as models which represent the amount of information an agent believes can be gained from selecting each arm. Even though they do not successfully predict period-by-period strategies, these models may still provide insight into how and why agents make the choices they do in bandits.

5.3.3 Model Conclusions

Of the five models considered, only ambiguity aversion and horizon truncation are consistent with the data. It is not surprising that horizon truncation is consistent because it is not restrictive, and generates no surprising predictions. Ambiguity aversion is appealing because it makes the surprising prediction that ambiguity averse agents will pay more than ambiguity neutral agents to learn the true mean of an ambiguous alternative.

In all likelihood, there is some merit to both of these models. It is unlikely that, without ambiguity aversion, such strong willingness to pay results could have been generated based on a "adjustment factor" alone. On the other hand, ambiguity

aversion cannot explain why information value ratios ever exceed one; constant, small errors in the adjustment factor could lead to increasing information value ratios as the optimal information value decreases.

5.3.4 Good News for Optimality

Although I have argued that the data strongly reject the optimal model, there are many important features of the data which could be interpreted as supporting optimality. In both the choice and Gittins index environments, subjects do experiment, and do acquire information about uncertain alternatives. In the choice framework, this leads to higher-valued arms being selected more frequently than lower-valued arms. In Chapter 4, Gittins indexes are higher for arms with higher expected values.

As information is acquired, the frequency of experimentation in the choice environment and the Gittins indexes decrease. The same holds true as the horizon varies. Indeed, in later periods in the choice experiment, the rate of experimentation in the data is not significantly different from that predicted by the optimal model. In Chapter 3, the Gittins indexes are very close to optimal after the first period.

Taken together, these results suggest subjects get right all the important comparative statics of the Gittins model, and are even pretty close to optimal in magnitude after the first period. However, there is still a significant problem with concluding that people behave optimally, and are maximizing their welfare. Even if they do well once they select an alternative one time, the evidence presented here suggests people behave far from optimally in determining whether or not to make the first choice of each alternative.

As was argued in Chapter 3, if agents do not appreciate the information value of the first selection of each alternative, they may not make the initial selection and will therefore never reach the domain on which they behave (near) optimally. If the untried alternatives do have high value, considerable welfare is lost. This problem looms particularly large in brand choice problems, where consumers who underestimate the information value of an initial trial may not try new products, depriving themselves

of potential welfare gains, and hampering the ability of the market to select new, improved products.

The models tested all preserve the comparative statics of the optimal model which are supported by the data, but also predict lower than optimal first period information values. In light of the results supporting optimality, the alternative models considered here could be considered an effort to understand the first period decision, for later-period optimality is little consolation if later periods are never reached.

5.4 Policy Implications

One important outcome of any economic research is policies which can help agents improve their welfare. This section considers cases where the models considered here can help shape policies designed to assist agents facing bandit problems, and some cases where policy is not necessary.

Before considering specific cases, however, it is worth asking whether helping people experiment more in bandit problems will do more harm than good. Encouraging experimentation might be harmful because, once they have some initial information, agents subsequently experiment too much. If the loss entailed in this overexperimentation is greater than that from underexperimentation, then the policy is harmful.

I argue that, in almost every case, overexperimentation is the more desirable outcome. This is true not only because experimentation is a public good, but also for another, more subtle reason. When overexperimenting, 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.

5.4.1 When Public Policy Can Help

In the naturally occurring environments in which ambiguity has been discussed, ambiguity aversion typically induces a pure taste for information. For instance, ambiguity aversion leads patients to request, and doctors to run, medical tests which will not affect their choice of treatment. In the multi-armed bandit environment, however, the ambiguity averse agent has another alternative: rather than incurring the cost of acquiring information to reduce the ambiguity of an ambiguous arm, she can simply select a less ambiguous arm. In most naturally occurring bandits, such an arm exists: the arm which has been tried before. This means that, without the opportunity to acquire information about ambiguous alternatives except by experimentation, ambiguity averse agents will simply avoid more ambiguous alternatives, and lock onto less ambiguous ones sooner than optimal.

This demand for information suggests policies which look a lot like those for risk aversion. Like risk averse agents, ambiguity averse agents will buy insurance. They would be willing to just ensure against bad probability outcomes, but since probability outcomes are unobservable, it is only possible to ensure against bad payoff outcomes, just as with risk aversion.

However, unlike risk averse agents, ambiguity averse agents are always willing to buy information about uncertain alternatives. In addition, their ambiguity aversion means the gain to be had from providing additional information is greater than that for ambiguity neutral agents. Therefore, there can be an additional social welfare gain by having government supply or subsidize the acquisition of information about ambiguous alternatives.

5.4.2 When Public Policy is Unnecessary

In many applications, public policy is unnecessary because other agents in the economy may have an interest in helping agents overcome their tendency to underexperiment. Because they are willing to pay more for information than ambiguity neutral agents, ambiguity averse agents provide an incentive for other agents to gather and

sell information about ambiguous alternatives.

There are many agents in the economy who fulfill this function. Consumer's Union tests all kinds of products, from orange juice to cars, and publishes *Consumer Reports* to provide information to agents facing brand choice problems. Although Consumer's Union is non-profit, for-profit niche magazines of all kinds conduct objective tests of products to provide information about the usefulness and value of competing products.

In addition to product testing organizations, surveying organizations, such as J.D. Power and Associates, reduce ambiguity by reporting the experiences of large samples of users of different products to those considering a purchase. Their statistics on automobile quality are widely published. This sort of information reduces ambiguity, rather than risk, because it uses a large sample to provide a good estimate of how likely a particular car is to be troublesome, but does not affect the probability of getting a defective car.

Often agents who stand to gain from additional experimentation can encourage information gathering. For instance, companies introducing new brands and stores with low prices would both like consumers to experiment with them. If these agents know consumers do not experiment enough, 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 tendency to underexperiment, 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 an oil company's 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 and 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.

5.4.3 When Public Policy is Needed Because of Underexperimentation

Unfortunately, while some agents have incentives to assist firms and consumers who undersearch and underexperiment, there are also incentives to exploit them. Agents who underexperiment 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 agents. Laws like the recent regulations requiring car dealers to clearly disclose down payment and financing information on leases can help consumers avoid situations where their tendency to underexperiment would lead them to compromise their future welfare.

5.5 Designing Replications and Extensions

Whenever experiments demand that subjects understand complex probabilistic structures, special care must be taken to ensure that they in fact understand how their payoffs are determined. I was fortunate in this research to have Caltech undergraduates as my subject pool. They are mathematically sophisticated enough to understand basic principles of probability and random sampling.

Past experiments on the Ellsberg paradox have asked subjects to understand compound lottery structures, but they have not included a time dimension, or asked subjects to do abstract evaluation like reporting a Gittins index or computing a willingness to pay for information. Even the analytically sophisticated Caltech subjects indicated that this problem was hard for them.

There are two factors which contribute to the complexity of these experiments, and they suggest ways in which other researchers working with different subject pools might try to reduce noise in their results. First, subjects may simply have difficulty understanding the problem itself. The combination of time structure and the compound lottery may not be clear to the subjects.

One way to help subjects understand the structure is to use actual probability generating devices. Rather than selecting payoff urns from a collection of actual urns containing actual balls, I used the balls and urns analogy to relate the probability structure. This allowed me to simulate an extremely large number of urns with an extremely large number of balls, which allowed me to use continuous approximations in my index and willingness to pay calculations. However, one could use a small number of urns and a small number of balls, which are actually available for subjects to examine and to help them understand the probability structure. Cox and Oaxaca (2000) use such a mechanism to convey a two-level probability structure to University of Arizona undergraduates with good results.

Another approach may be to add context to the experimental instructions. Although experimentalists are generally wary of using context for fear of uncontrolled effects, adding context to these experiments may use subjects' understanding of certain naturally occurring circumstances to help them understand the information and incentive structure of the problem. I actually developed a set of instructions for the experiment in Chapter 3 built around choosing to commute by car or train.

Experience also enhances understanding of the environment. Subjects were given two guided practice bandits in each experiment. Although several people still had questions after the instructions were read aloud, being walked through two sample bandits alleviated most of their confusion. Although other less analytically sophisti-

cated subject pools might require more experience, using the mechanism undoubtedly helps them understand it.

The second factor which makes these experiments difficult for subjects is, once they understand how their payoffs are determined, actually determining the best strategy. Pilot experiments suggested, especially in when eliciting the Gittins index, that it was not obvious to subjects that the Gittins index should be higher than the expected value of an arm. Since my objective was to understand whether or not information values were close to optimal given a subject understood they exist, I edited the instructions to include some strategic advice explaining why information values will be positive, and selected the guided practice periods to demonstrate that the information value should be positive.[1] My impression was that the additional instructions helped subjects a little, but the guided practice periods really helped subjects understand how the information value arises.

Given that subjects understand that there is an information value, another difficult aspect of determining the right strategy is computing the magnitude of the information value. Even Caltech subjects did not manage to accurately compute Gittins indexes, though they managed to do so with systematic biases; whether or not subjects computed optimal indexes is what the experiment was designed to test. It is probably reasonable to expect that less analytically sophisticated subjects will not be better at solving the dynamic programming problem necessary to compute the Gittins index.

Nevertheless, other subject pools may benefit from some training to help them understand comparative statics which were intuitive to Caltech subjects. Dedicated training sessions have been used to teach, through simple examples, the basic solution techniques to problems posed in subsequent experiments. However, it should be noted that while Caltech subjects understood some simple statics, like the information value gets smaller as additional information is obtained, they did not understand other features, like the effect of the horizon or prior mean in Chapter 4.

[1] My impression is that figuring out the information value should be positive was more a sudden epiphany than a slow learning process. For this reason, I did not consider it to be interesting in itself.

5.6 Avenues for Future Research

This set of experiments has expanded economists' understanding of the factors which influence behavior in economically significant multi-armed bandit problems. However, it has also raised a number of important questions which may lead to further improvements of our understanding of bandits, and suggest additional ways in which agents in bandit problems might be helped to improve their welfare.

5.6.1 Choice

As discussed in Chapter 2, actually computing the indexes for an interesting bandit with equally risky arms is computationally challenging. With technological improvements, it may be possible to perform these computations for certain clever experimental designs. For instance, if the sequence of payoffs from each arm were known to the experimenter beforehand, the number of indexes which would need to be computed would be dramatically reduced. Although each would be time consuming, the difference may be enough to develop a satisfactory notion of optimality for a normal arm with an informative prior. This would allow analysis of choice data from an experimental bandit with a well-controlled prior.

In addition to computing a better approximation to the optimal strategy, choice data may be improved by expanding the set of choices. In the bandits presented in Chapter 2, there were enough periods and few enough bandits that sampling could be conducted through cyclical algorithms. If there were more choices than the expected number of periods, however, subjects would be forced to pay attention to initial payoffs, and more carefully consider when sampling was appropriate.

5.6.2 Solving Bandits

Horizon truncation models how agents reduce a complex dynamic programming problem to a simple one. One way to test this model against ambiguity aversion would be to have subjects play a very simple bandit which could be easily solved explicitly.

For instance, the example given in Chapter 1 can be solved with a few quick algebraic steps. An agent would not need to truncate the horizon to solve that problem. However, if she were ambiguity averse, she would still have lower than optimal Gittins indexes.

Similarly, subjects might be given a training session which taught them about dynamic programming. Although the strategy section of the instructions in Chapters 3 and 4 discussed the factors which lead to variation in information values, a more thorough explanation might alter how agents approached these problems. If this resolved the suboptimalities, it would suggest education as an important aspect of any policy directed at people facing bandit problems.

5.6.3 Bandits in the Market

One particularly economically significant form of experimentation cost is that imposed by markets: prices. Grether et al. (1988) found support for three models of price search in their laboratory experiments. In a brand choice environment, arms are goods which yield different payoffs. The distribution of payoffs arises from variation in product quality, and from individual variations in tastes. However, if one good is consistently better, its price will rise, and send a signal to others who have not tried it.

Since so much commerce takes place in such markets, it is worthwhile to study how the addition of a market structure affects behavior in bandits. If all agents have the same preferences, prices may send a signal. In this case, experimentation may be a public good, which helps even those who did not initially experiment with the best alternative. If agents have different preferences, so market prices contain little information, the market may reduce to the individual choice case.

Since preferences are often only correlated, and differences in prices will affect net surpluses, most naturally occurring markets may fall between these two extremes. Understanding how a market structure affects bandit strategies may help develop a crucial link between individual behavior, and individual deviation from optimality,

and aggregate behavior and disequilibrium.

5.7 Concluding Remarks

Although in detail this study examined behavior in multi-armed bandit problems, the broader goals accomplished by the method deserve some attention. Chapters 3 and 4 develop formal models of how decision makers who do not satisfy all the typical assumptions of rationality might behave in bandit problems. The primary point of Chapter 3 is that hyperbolic discounters' Gittins indexes will be lower than optimal, by a constant factor, in all periods, but experimental subjects' indexes are increasing; the primary point of Chapter 4 is that ambiguity averse agents' Gittins indexes will be lower than optimal but their willingness to pay will be higher than optimal, and that this is true for experimental subjects. However, the theoretical exposition in each chapter is built on the subtext that it is possible to use simple behavioral models to make predictions about complex economic problems.

The extension of the basic set of bandit results to hyperbolic discounting is straightforward because, for the naive hyperbolic discounters considered here, time inconsistency is not a factor in the computation of the index. The extension to ambiguity aversion is much less straightforward, but nonetheless possible. The theoretical exercise is worthwhile because these extensions generate predictions which can be used to test for the presence of these phenomena in simple experiments.

Given the complexity of computing Gittins indexes, and the restrictions and simplifications necessary to even be able to numerically determine them, one might argue it is not surprising that experimental subjects are unable to calculate optimal Gittins indexes. Nevertheless it is valuable to understand exactly how agents deviate from optimality, and why.

Although subjects clearly do not use Bayes rule to determine their beliefs in every possible future state, use backward induction to determine the expected value of a given strategy and policy improvement algorithms to determine the best strategy,

they nevertheless have a sense of how much they might learn from experimenting with each arm. Computing the exact numerical solution is very complex, but getting close, and having the right comparative statics, could be quite intuitive. The utility of the optimal model, and the alternatives considered here, is to provide economists with a formal representation of how people behave; the models considered here are useful as long as agents behave as if they are using the models. In addition, the optimal model provides a normative benchmark against which behavior can be measured to determine if policies can help agents improve their welfare.

What sort of mechanisms might lead agents to behave as if they are using an optimal model? They may solve a smaller version of the problem, as suggested by horizon truncation, which would generate comparative statics which are similar to those in generated by the optimal model. Similarly, a hyperbolic, risk averse, ambiguity averse or quantal response agent might solve a smaller version of the problem and generate comparative statics which resemble those of the model which generated them.

Agents may also have simple rules which the use to understand the value of information. For instance, they may begin with a baseline information value, and decrease it linearly or exponentially as additional information is obtained. Since a linear decrease is not a bad approximation after the first period, the quality of this rule depends on how the initial value is determined. One possibility is that agents use the same initial information value for every bandit. For individuals who encounter a broad range of bandits, with different payoff scales and variances, this will not in general lead to optimal or close to optimal strategies. However, agents repeatedly encounter bandits with similar scales and variances, as might oil exploration firms or pharmaceutical researchers, may have an initial information value which is well-calibrated from experience, and which could provide very close to optimal behavior.

However, the evidence here suggests that agents are sensitive to changes in the scale and variance of their payoffs, as information values change in response to prior beliefs. This suggests that if agents are using a linear decrease rule, they are determining their initial information values individually for each bandit. The models

considered are designed to help us understand that process, and ultimately guide us in how that process might be improved.

In choice environments, it is not necessary for agents to formulate an initial information value, or to quantify an intuitive sense of the need to further experimentation. Instead, agents could use cycling rules like those discussed in Section 2.6.2. If agents do not cycle enough, they will appear to underexperiment. If they have a sense that this leads to underexperimentation, they may compensate by experimenting more later.

The fact that different mechanisms may operate in choice and Gittins index frameworks suggests that there may be a crucial disconnect, not considered here, between the way agents solve bandit (choice) problems and the way they compute Gittins indexes. Although the choice data are broadly consistent with the results about Gittins indexes presented here, it is important to understand they may arise from fundamentally different processes which may or may not produce similar results in any given problem. Further research is necessary to understand the fungibility of choice and Gittins index results.

One of economists' most important roles is to contribute to the public policies which help people improve their welfare and to corporate strategies which help companies maximize profits in the face of individual suboptimalities. Understanding why people do not behave optimally provides important insights into which policies and strategies will be effective and which will not. This understanding is a product of developing and testing models based on behavioral assumptions.

Citations

Abramowitz, M. and C. Stegun. (Eds.). Confluent Hypergeometric Functions. Ch. 13 in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing. New York: Dover, 1972. 503-515.

Anscombe, F. and R. Aumann. A Definition of Subjective Probability. *Annals of Mathematical Statistics* 34, 1963. 199-205.

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.

Blume, A. D. DeJong, G. Neumann and G. Savin. Inferring Learning Rules from Experimental Game Data. Mimeo. 1999.

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. and T. Ho. Experience-Weighted Attraction Learning in Normal Form Games. *Econometrica* 67:827-874. 1999.

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. Good News and Bad News: Search from Unknown Wage Offer Distributions. *Experimental Economics* 2:197-226, 2000.

——. 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.

de Finetti, B. Probabilities of Probabilities. In *New Directions in the Application of Bayesian Methods.* Ed. A. Aykac and C. Brumat. Amsterdam: North Holland, 1977. 1-10.

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

Einhorn, H. and R. Hogarth. Ambiguity and Uncertainty in Probabilistic Inference. *Psychology Review* 92:433-461, 1985.

Ellsberg, D. Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics* 75:643-669, 1961.

Frisch, D. and J. Baron. Ambiguity and Rationality. *Journal of Behavioral Decision Making* 1:149-157, 1988.

Gilboa, I. and D. Schmeidler. Maxmin Expected Utility with a Non-Unique Prior. *Journal of Mathematical Economics* 18:141-153, 1989.

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.

Grether, D., A. Schwartz and L. Wilde. Uncertainty and Shopping Behavior: an Experimental Study. *Review of Economic Studies* 60:323-342. 1988.

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.

Hogarth, R. and H. Einhorn. Venture Theory: A Model of Decision Weights. *Management Science* 36:780-803, 1990.

Kahn, B. and R. Sarin. Modelling Ambiguity in Decisions Under Uncertainty. *Journal of Consumer Research* 15:265-272, 1988.

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

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

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

McFadden, D. Modeling the Choice of Residential Location. In *Spatial Interaction Theory and Residential Location.* Ed. A. Karlquist et al. Amsterdam: North Holland, 1978. 75-96.

McKelvey, R. and T. Palfrey. Quantal Response Equilibria for Extensive Form Games. *Experimental Economics* 1:9-42. 1998.

——. Quantal Response Equilibria for Normal Form Games. *Games and Economic*

Behavior 10:6-38. 1995.

——. 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.

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

Quiggan, J. A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization* 3:323-343, 1982.

Rubinstein, A. Similarity and Decision-making Under Risk (Is There are Utility Theory Resolution to the Allais Paradox?). *Journal of Economic Theory* 46:145-153, 1988.

——. Is it "Economics and Psychology"?: The Case of Hyperbolic Discounting. Mimeo. 2000.

Salmon, T. An Evaluation of Econometric Models of Adaptive Learning. Mimeo. 1998.

Schmeidler, D. Subjective Probability and Expected Utility without Additivity. *Econometrica* 57:571-587, 1989.

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

Segal, U. The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach. *International Economic Review* 28:175-202, 1987.

Siegel, S. and N. Castellan. *Nonparametric Statistics for the Behavioral Sciences*, 2 ed. New York, NY: McGraw-Hill Book Company, 1988.

Slater, L. *Confluent Hypergeometric Functions.* Cambridge: Cambridge University Press, 1960.

Spanier, J. and K. Oldham. The Kummer Function $M(a, c; x)$. Ch. 47 in *An Atlas of Functions.* Washington, DC: Hemisphere, 1987. 459-469.

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

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

Wolfram Research. The Fisher-Tippett Distribution. In *Eric Weisstein's World of Mathematics.* http://mathworld.wolfram.com/Fisher-TippettDistribution.html. Accessed 9/6/00.

Wolfram Research. Confluent Hypergeometric Function of the First Kind. In *Eric*

Weisstein's World of Mathematics. http://mathworld.wolfram.com/ ConfluentHypergeometricFunctionoftheFirstKind.html. Accessed 5/31/00.

Yates, F. and L. Zukowski. Characterization of Ambiguity in Decision Making. *Behavioral Science* 21:19-25, 1976.