Theory Dec.
DOI 10.1007/s11238-011-9259-2

Ambiguity aversion in multi-armed bandit problems

Christopher M. Anderson

Abstract In multi-armed bandit problems, information acquired from experimentation is valuable because it tells the agent whether to select a particular option again in the future. This article tests whether people undervalue this information because they are ambiguity averse, or have a distaste for uncertainty about the average quality of each alternative. It is shown that ambiguity averse agents have lower than optimal Gittins indexes, appearing to undervalue information from experimentation, but are willing to pay more than ambiguity neutral agents to learn the true mean of the payoff distribution, appearing to overvalue objectively given information. This prediction is tested with a laboratory experiment that elicits a Gittins index and a willingness to pay on six two-armed bandits. Consistent with the predictions of ambiguity aversion, the Gittins indexes are significantly lower than optimal and willingnesses to pay are significantly higher than optimal.

Keywords Bandit problem · Gittins index · Ambiguity aversion

JEL Classification C91 · C44 · C73 · D83

1 Introduction

In many economically significant environments, agents must repeatedly choose among uncertain alternatives about which they can learn only through experimentation. Individuals face problems such as brand choice, where shoppers must decide whether to purchase a favorite brand of orange juice or experiment with a new one, and job or

C. M. Anderson (✉)
Department of Environmental and Natural Resource Economics, University of Rhode Island, Kingston, RI 02881, USA
e-mail: cma@uri.edu

Published online: 19 June 2011

price search where searchers must decide whether to accept a current option, or to continue searching for a better offer. Project managers must use successive phases of experimentation to decide when to abandon research on approaches to problems that are unlikely to produce a breakthrough, and devote resources to the most promising approaches. Oil companies must decide when to discontinue testing a mine or oil lease, and either commit to mining or drilling it, or not. 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 every day 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.

The information structure in these examples can be modeled as a multi-armed bandit problem.[1] Bandit problems present agents with a choice among alternatives (or arms) of unknown quality, so experimentation is required to identify the best option. However, trying an alternative is not perfectly revealing, as average quality is observed with noise. For instance, a consumer may get a bad carton of orange juice from a good brand, or a single test might hit sweet crude on an otherwise dry tract. Through repeated sampling, agents use draws from the distribution of payoffs of an arm to learn about its average quality; Bayes' rule is used to update second-order beliefs about the first-order distribution of payoffs from an alternative. Since there is an opportunity cost to experimentation, at each decision agents are faced with the dilemma between exploiting the alternative which has performed best in the past and experimenting with less known alternatives which they may discover to be better. If an agent finds the less familiar arm to be better than her initial beliefs suggested, she can exploit it in the future to increase her payoffs, but if it is not, she can switch back to the original good performer. Optimal play depends on agents correctly assessing the value of information, in the form of increased expected payoffs, arising from experimenting with the less known alternatives.

Despite the range and economic importance of problems that can be represented as multi-armed bandits, how agents approach such problems is only beginning to be explored. One group of experiments focuses on whether (rather than precisely how much) subjects value information acquired through experimentation. Banks et al. (1997) present subjects with a two-armed bandit in which the expected payoff of one arm is known and the other is not. They find experimentation is more common in a treatment where initial selection of the unknown arm is optimal than in a treatment where experimentation is not optimal, suggesting people do understand there is value to information obtained through experimentation. Norton and Isaac (2010) extend this result to a principal-agent environment where the ambiguity is induced by strategic uncertainty, rather than multi-level lotteries. Boyce et al. (2009) find information is generated (although suboptimally) in environments with an information spillover, but the effects of undervaluation and free ridership cannot be easily separated.

Another experimental approach to bandits attempts to identify patterns in choices that are consistent with a catalog of ad hoc decision rules. Meyer and Shi (1995)

[1] The name "bandit problem" refers to the canonical formulation where an agent must decide how to allocate a number of pulls among a set of slot machines which pay with different, unknown probabilities.

consider two-armed bandits with one arm's mean known, and have difficulty identifying a unifying explanation among risk and ambiguity attitudes and other decision rules; shortened planning horizons and simple stick-with-a-winner logic fit the data better than the considered alternatives; Acuna and Schrater (2008) find similar limited-memory Bayesian models are best performing in 3- and 4-arm bandits. Gans et al. (2007) consider a broad range of ad hoc models in a two-armed bandit, and find backward-looking strategies which predict switching arms after a fixed number of consecutive failures best explain the data. In these studies, the optimal choice model did not fare well. Steyvers et al. (2009) associate use of different heuristic decision rules with a range of psychological and demographic measures; optimal play is correlated with higher IQs, but little else.

Collectively, this work presents convincing evidence that people do appreciate that there is value to experimentation, and suggests that they systematically underestimate it. This is problematic because optimal strategies trade off differences among arms' expected values based on current beliefs, and differences in the information on which those beliefs are based. Therefore, even if an agent undervalues the information to be acquired by experimenting with each arm by, say, a factor of one-half, she will not be able to properly trade off differences in information with differences in expected value.[2]

This undervaluation is consistent with evidence from closely related dynamic search problems, undertaken to test models of risk and ambiguity attitudes, rather than decision rules. Pratt et al. (1979) find that empirical price dispersion for a range of goods is higher than could be supported by optimal price search, leading to less competitive markets. In laboratory experiments on wage offer search, Cox and Oaxaca (1989, 1990 , 1992, 1996, 2000) and Schotter and Braunstein (1981) find that subjects do not search enough, suggesting an undervaluation of payoff opportunities resulting from continued search, even when offer recall is possible (Cox and Oaxaca 1996). In a treatment paralleling a bandit problem, undersearch is more severe when it is not known which of two distributions is generating offers (Cox and Oaxaca 2000). However, the authors reject search cost as a potential explanation, and the level of global risk aversion implied by the data is preposterous. Without an understanding of why agents' bandit strategies are suboptimal, it is difficult to develop carefully tailored public policies and marketing strategies to help improve the welfare of those facing bandit problems.

Rather than attempt to deduce whether any ad hoc rule is being used based on choice data, this article adopts a different approach to understanding behavior in bandit problems. It derives and tests directly a surprising prediction of a single theory, the optimal model adapted to account for ambiguity aversion. Ambiguity distinguishes bandit problems from lottery choice problems because the distribution of possible payoffs from each arm is unknown; there is a second-order belief over possible values of *average* quality. Experimental evidence from static choice problems suggests that

[2] An extreme case is one where a consumer has a favorite brand, which she has been buying for so long she has a very good idea of its expected value. When a new brand is introduced, her scaled-down information value will guide her to stick with her favorite brand if her favorite brand is enough better than her beliefs about the new brand, even when experimentation would be optimal.

people dislike spread in the second order distribution, ambiguity aversion (Ellsberg 1961; Hogarth and Einhorn 1990; see Camerer and Weber 1992 for a survey). In the dynamic bandit setting, aversion to ambiguity may cause people to select alternatives with which they are more familiar, a cause for observed underexperimentation which may require different remedial measures than those for simple myopia.

In formally incorporating ambiguity aversion in the optimal choice model for bandit problems, a seemingly paradoxical prediction is generated: agents will be willing to accept a lower-than-optimal known expected value alternative in exchange for an uncertain alternative, 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 contrasts with predictions of shortened planning horizon models, risk aversion, and backward-looking stick-with-a-winner or switch-from-a-loser heuristics.[3] This prediction is tested in an experiment, which provides strong support for this surprising prediction of ambiguity aversion. The final section discusses the implications of ambiguity aversion for buyers and sellers in private markets, as well as public policy.

2 The bandit problem

In a bandit problem, the agent's objective is to select a sequence of arms that maximizes the present value of the sequence of payoffs received. The bandits considered here have two arms that yield Bernoulli distributed payoffs, a 1 in the event of a success and a 0 in the event of a failure. The "known" arm has a known probability, λ, of yielding a success, and a $(1-\lambda)$ probability of failure. The other, ambiguous arm has an unknown probability of success, θ, where θ is drawn from a known distribution, with cumulative density G and probability density g. The agent has beliefs, F, about the cumulative distribution of θ, with corresponding probability density function f. The agent's prior is that $F = G$, but F is then updated using Bayes' rule as the arm is selected and payoffs are observed. The agent is permitted n total arm selections, and her total payoff is the sum of her n payoffs.[4] The bandit is then denoted $(F, \lambda; n)$.

2.1 Optimal bandit strategy: the Gittins index

A theoretical and empirical benchmark for choices in the bandit problem is the Gittins index, $\Lambda(F, n)$. The Gittins index is the value of a known mean (i.e., known λ) arm for which the agent is indifferent between selecting an unknown arm with beliefs F and the known-mean arm in the $(F, \Lambda(F, n); n)$ bandit in the current period. This formulation reduces the bandit to an optimal stopping problem, since once the known-mean arm is chosen, no updating on the unknown-mean arm occurs that could support switching

[3] It is difficult to analytically compare predictions of forward-looking optimization models with history-dependent heuristics which make decisions based on observed outcomes, rather than possible future outcomes.

[4] The main results presented extend to bandits with geometric discounting, and other discount sequences with the stopping property (Anderson 2001).

back. Conceptually, an arm's index is the sum of its expected value given current beliefs, $E[\theta|F]$, and its information value. The information value arises from the possibility that one could learn, though subsequent experimentation, that the true payoff probability is better than $E[\theta|F]$, and exploit this knowledge in all future periods; if experimentation reveals the payoff probability to be worse than $E[\theta|F]$, the agent can choose the known $\Lambda(F, n)$. Gittins and Jones (1974) show that the optimal strategy in a bandit with K unknown arms is to select the arm with the highest Gittins index.

2.2 Bandits with ambiguity averse agents

Ambiguity arises in bandits through the second-order belief about payoffs, F (Anscombe and Aumann 1963).[5] Aversion to second-order uncertainty about the payoff distribution was first captured in the Ellsberg paradox (1961) and has since been modeled with subadditive probabilities (e.g., Schmeidler 1989), multiple probabilities with minimax-style decision rules (e.g., Gilboa and Schmeidler 1989) and by relaxing the assumption of reduction of compound lotteries between the first- and second-order beliefs (e.g., Segal 1987; Kahn and Sarin 1988). (See Camerer and Weber (1992) for a survey.)

Since bandit problems have explicit second order probabilities in beliefs F over payoff probabilities θ, these models are the most natural to apply. Segal (1987) and Kahn and Sarin (1988) propose decision weighting functions, $\omega(F)$, that agents use in evaluating ambiguous lotteries. They incorporate ambiguity attitudes by not requiring the decision weight assigned to a second-order lottery be $\omega(F) = E[\theta|F] = \int_0^1 \theta f(\theta) d\theta$, as reduction of compound lotteries suggests. Instead they reflect distaste for ambiguity by shifting subjective probability weight to lower values of θ than the objective second order probability F indicates, expressing ambiguity aversion as pessimism about the true payoff probability. The decision weight is then used, in place of the objective probability of payoff $E[\theta|F]$, to calculate an analog of expected utility and rank available alternatives. Ambiguity aversion occurs when the decision-weighted utility of an ambiguous choice is less than its objective expected utility.

Definition 1 Suppose F is an agent's belief about the payoff probability θ of a Bernoulli lottery. An agent with decision weighting function $\omega(F)$ is *ambiguity averse* if $E[\theta|F] - \omega(F)$ is positive and monotonically decreasing as the variance of F decreases, with $\omega(F) = E[\theta|F]$ when the variance of F is zero.

That is, the agent is ambiguity averse if the decision weight used to evaluate an ambiguous alternative is less than its expected value, but the discount relative to objective probability gets smaller as the agent becomes more certain about the value of θ. This

[5] No consensus has emerged around a single definition of ambiguity. 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 that 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 (de Finetti 1977).

Bernoulli analysis assumes the utilities from a success and a failure to be 1 and 0 respectively, allowing direct comparison of $E[\theta|F]$ with the decision weight $\omega(F)$.[6]

A challenge in adapting these static notions of choice under ambiguity to the bandit problem is identifying how information acquired through sampling is incorporated. In bandit problems, ambiguity neutral agents use Bayes' rule both to update prior beliefs to reflect new payoff observations and to reduce the compound lottery between F and θ. New models by Klabinoff et al. (2009) and Traeger (2010) allow uncertainty at different levels of compound lotteries to be discounted differently in dynamic environments. In the model proposed here, it is assumed that all agents understand and use Bayes' rule to update prior beliefs, F, but do not reduce compound lotteries because of their distaste for ambiguity, rather than a misunderstanding of how to do so with Bayes' rule.[7] This is similar to how a risk averse agent may understand how to calculate an expected value, but nevertheless willingly pay a risk premium.[8]

To emphasize the distinction between Bayesian updating of objective future states and the choice of an ambiguous alternative, objective payoff probabilities will be denoted $E[\theta|F]$, and are not affected by the agents' ambiguity attitudes. F_1 denotes that F has been updated using Bayes' rule to reflect a success, and F_0 a failure. To indicate that a given arm is being evaluated by an ambiguity averse agent, the evaluation function will be written with a superscript ω.

The next two sections apply ambiguity aversion to derive the primary theoretical results, that the revealed value of information about θ depends on the frame in which it is elicited. The first result, from the arm choice frame, shows that ambiguity averse agents have lower Gittins indexes than do ambiguity neutral agents, thus appearing to value less the information obtained from experimentation, and accordingly pursue it less aggressively. However, the following section, from the information purchase frame, shows that ambiguity averse agents are willing to pay more than their ambiguity neutral counterparts to learn the payoff probability of the ambiguous arm.

[6] Risk neutrality is assumed because introducing a second source of payoff function curvature confounds general results about the effects of ambiguity aversion. By specifying particular parametric forms for risk and ambiguity aversion, it would be possible to identify regions of the parameter space where Theorems 1 and 2 still hold. The experiments are designed so that results can still be interpreted in the presence of risk aversion; see footnote 12.

[7] This approach assumes agents have a mental model where the ambiguity involved in arm choice is resolved each after selection. An alternative mental model would view a bandit with horizon n as a n-level compound lottery over final payoff states, and compute value by reducing the compound lottery (or not) to determine the probability of each state. This would imply a different analytical structure, but is a sensible alternative for future development and testing.

[8] In a static choice experiment designed to distinguish ambiguity attitudes from the ability to reduce compound lotteries, Halevy (2007) found ambiguity neutrality was common among the 30% of subjects who understood reduction of compound lotteries, and ambiguity aversion was common among those who violated reduction of compound lotteries. However, his experiment did not address the question of whether subjects view sequences of compound lotteries as individually resolved or a large compound lottery over final payoff states.

2.3 The Gittins index of an ambiguity averse agent

Since they have a distaste for ambiguity, ambiguity averse agents are less willing to experiment with more ambiguous alternatives. This corresponds to having a smaller Gittins index than ambiguity neutral agents.

Theorem 1 *Suppose $\omega(F)$ is an ambiguity averse decision weighting function. Then* $\Lambda(F, n) \geq \Lambda^{\omega}(F, n)$.

This proof is by contradiction. Suppose $\Lambda(F, n) < \Lambda^{\omega}(F, n)$. Then ambiguous arm 1 is optimal initially in the $(F, \Lambda(F, n); n)$ for both ambiguity neutral and ambiguity averse agents. Let $V(F, \lambda; n)$ denote the value received from the sequence of arm choices that maximize the agent's objective function in the $(F, \lambda; n)$ bandit. Then

$$\begin{aligned} V(F, \Lambda(F, n); n) - V^{\omega}(F, \Lambda(F, n); n) &= E[\theta|F] - \omega(F) + E[\theta|F] \\ &\quad \left[V(F_1, \Lambda(F, n); n-1) - V^{\omega}(F_1, \Lambda(F, n); n-1)\right] \\ &\quad + (1 - E[\theta|F]) \left[V(F_0, \Lambda(F, n); n-1) - V^{\omega}(F_0, \Lambda(F, n); n-1)\right]. \quad (1) \end{aligned}$$

Note that the ambiguity neutral and ambiguity averse agents view the expected payoff probabilities $E[\theta|F]$ in the third and fourth terms objectively, and thus they are not transformed by the ambiguity function. After making her choice, the ambiguity averse agent anticipates learning the payoff outcome and correctly applies Bayes' rule to update her beliefs F in the event of a success or failure, and is able to correctly apply F to calculate the probabilities of each continuation. She understands these probabilities, but does not like the uncertainty they imply for her payoffs.

The first term on the righthand side of Eq. 1 is nonnegative by the definition of ambiguity aversion. The second and third terms are each positive because the value of the ambiguity discounted bandit will necessarily be less than the value of the same bandit evaluated by an ambiguity neutral agent.[9] Thus, the difference in Eq. 1 is positive, so ambiguity averse agents ascribe lower values to bandits than do ambiguity neutral agents.

However, $\Lambda(F, n)$ being the Gittins index implies that using known arm 2 at every stage is optimal for the ambiguity neutral agent in the $(F, \Lambda(F, n); n)$ bandit. Therefore

$$\begin{aligned} V(F, \Lambda(F, n); n) - V^{\omega}(F, \Lambda(F, n); n) &= n\Lambda(F, n) - V^{\omega}(F, \Lambda(F, n); n) \\ &\leq 0 \quad (2) \end{aligned}$$

The inequality follows because $\Lambda(F, n) < \Lambda^{\omega}(F, n)$ implies that known arm 2 cannot be the optimal choice for the ambiguity averse agent in the first period of the $(F, \Lambda(F, n); n)$ bandit. Therefore, the value calculated by the ambiguity averse agent must be strictly greater than $n\Lambda(F, n)$ (arrived at through some strategy which selects unknown arm 1 initially). However, the difference expressed in Eq. 2 being negative contradicts the difference in Eq. 1 being positive. Therefore, $\Lambda(F, n) \geq \Lambda^{\omega}(F, n)$,

[9] A detailed proof is available from the author.

with strict inequality if it is optimal for the ambiguity neutral agent to initially select the ambiguous arm. □

The index of an ambiguity averse agent must be interpreted carefully. Theorem 1 shows that if an ambiguity averse agent calculates an index using her utility function and the optimal algorithm for an ambiguity neutral agent, she gets a lower value. If she plays an index strategy, she will thus underexperiment, and receive a lower payoff. While it is intuitive that ambiguity aversion will bias agents away from unfamiliar alternatives, it has not been shown that an index strategy is optimal (in some sense) for an ambiguity averse agent.

2.4 Willingness to pay for information about the true average payoff of an ambiguous arm

The next theorem shows that ambiguity averse agents are willing to pay more than ambiguity neutral agents to learn the value of θ. This "ambiguity premium" is analogous to a risk premium, additional money agents are willing to pay to avoid uncertainty they dislike.

Theorem 2 *Let $\Delta(F, \lambda; n)$ denote the maximum amount the agent is willing to pay to learn the true value of θ in the $(F, \lambda; n)$ bandit. Then $\Delta(F, \lambda; n) \leq \Delta^{\omega}(F, \lambda; n)$, with equality only when $\lambda \geq \Lambda(F, n)$.*

Proof If the θ is known, the bandit is not ambiguous and both the ambiguity averse and ambiguity neutral agent will select the higher expected value arm on all trials. They value this lottery choice problem using their beliefs F to expect over possible values of θ:

$$n\left[\int_0^{\lambda} \lambda f(\theta)\mathrm{d}\theta + \int_{\lambda}^{1} \theta f(\theta)\mathrm{d}\theta\right] \tag{3}$$

The ambiguity averse agent is able to calculate the expected value of knowing θ without accounting for ambiguity because the realization of θ is resolved prior to the determination of payoffs, and thus there is no second order probability.

$\Delta(F, \lambda; n)$ is the difference between Expression 3 and $V(F, \lambda; n)$ for the ambiguity neutral agent; $\Delta^{\omega}(F, \lambda; n)$ is the difference between Expression 3 and $V^{\omega}(F, \lambda; n)$ for the ambiguity averse agent. Since Expression 3 is the same for both, the effect of ambiguity aversion is only in the difference between $V^{\omega}(F, \lambda; n)$ and $V(F, \lambda; n)$. An argument parallel to that following Eq. 1, with λ replacing $\Lambda(F, n)$, shows that $V^{\omega}(F, \lambda; n) \leq V(F, \lambda; n)$. Equality occurs only when $\lambda \geq \Lambda(F, n)$, and both agents select the known arm in the ambiguous bandit. Thus, the willingness to pay of the ambiguity averse agent to eliminate ambiguity from the choice problem is greater than that of the ambiguity neutral agent. □

This theorem provides a surprising contrast to Theorem 1. Because ambiguity averse agents are willing to pay more than ambiguity neutral agents to learn the true value

Table 1 Properties of the six unknown arms used in the experiment

	Prior beta (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

of θ, it appears that they overvalue information. On the other hand, when the information value question is presented in the arm choice frame, because their Gittins index is lower than optimal, ambiguity averse agents appear to undervalue information.

This paradox does not arise in other models that have been used to understand bandit behavior. Ambiguity averse agents value the counterfactual universe where they know θ without ambiguity aversion; agents who are risk averse, backward-looking (e.g., Gans et al. 2007) or who plan over a shortened horizon (e.g., Meyer and Shi 1995), will bring these suboptimalities to the counterfactual calculation. Therefore, this surprising prediction of ambiguity aversion is excellent grounds for testing the model.

3 Experimental design

To test the model of ambiguity aversion in multi-armed bandits, an experiment is designed in which subjects play the same six two-armed, one arm known, Bernoulli bandits in each of two treatments. In one treatment, a mechanism is used to elicit a Gittins index for the unknown arm of each bandit problem. In the other treatment, a maximum willingness to pay for perfect information about the unknown mean arm is elicited. The elicited Gittins indexes and willingnesses to pay can then be compared to the ambiguity neutral predictions to test for the effects of ambiguity aversion demonstrated in Theorems 1 and 2.

Subjects in different sessions encountered the treatments in different orders, and within each treatment, they played the six Bernoulli bandits in random order. Table 1 shows the parameters of the six unknown mean (ambiguous) Bernoulli arms used in the experiment. The value of θ, the probability of success, is is distributed beta(α, β), where α and β are known parameters. In each bandit, the ambiguous arm is paired with a known-mean (unambiguous, but still risky) arm, whose characteristics vary with treatment and are described below, to form a two-armed bandit. It is the Gittins index of, and willingness to pay for, information about the unknown mean arm that are of primary interest.

The payoff probabilities are explained 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. Each period, a ball is drawn at random from the chosen urn and a payment of 1 is made if it 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. Since the emphasis in the experiment is on information value rather than updating, the experimental software helps subjects by providing them with a "best guess" at the number of red balls in the unknown urn, which "was arrived at using a law of probability called Bayes' rule"[10]

The instructions used in each treatment are available from the author.

3.1 Gittins index treatment

In the Gittins index treatment, each of the six bandits begins by telling the subject the prior for the unknown mean arm and number of periods in the bandit, but only that the mean of the known mean arm will be determined randomly. She is then asked a series of questions of the form, "Would you choose the Known [mean] urn this period if it had Z red balls?" The initial Z is randomly chosen from [0, 100] to control for anchoring. The question is repeated, with successive values of Z given by a bisection algorithm, until the subject's indifference point is narrowed to the nearest percent, her Gittins index for the unknown mean arm. The mean of the known mean arm is then announced, and the subject's first period arm choice is made for her 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.[11] In subsequent periods, the subject can choose either the known or unknown mean arm.

3.2 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 0.5 in each of the six bandits. Before the first period, the subject is asked a series of questions of the form, "Would you pay \$Z to learn the number of red balls in the unknown [mean] urn?" This question is repeated, with successive values of \$Z (between −\$0.05 and \$1.00) given by a bisection algorithm, until the subject's indifference point is narrowed to the nearest penny, her 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. She may then choose either arm.

In this treatment, the value of learning θ depends on the value of the known mean arm, λ. Theorem 2 proves that for any fixed $\lambda < \Lambda(F, n)$, the ambiguity averse agent will pay strictly more than an ambiguity neutral agent to learn the value of θ; if $\lambda \geq \Lambda(F, n)$, then both agents will pay the same. The design uses $\lambda = 0.5$, so the

[10] A few subjects explicitly rejected the best guess, claiming past payoff realizations provided a better estimate. This suggests neglect of base rates (Camerer 1995) may be more than just a cognitive shortcut.

[11] Note that because her first period arm choice is made based on her reported value, this mechanism is incentive compatible, and truthfully elicits the Gittins index regardless of the distribution from which the mean of the known mean arm is chosen.

different unknown arm 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.[12]

4 Data and results

Data was collected on 33 Caltech undergraduates,[13] who did not necessarily have any training in economics, though many had participated in unrelated economics experiments. Subjects were paid for every dollar (red ball) they earned in excess of $28, leading to average payments of $18 (range $3 to $24) for sessions lasting about 90 minutes.

Before directly testing the predictions of the ambiguity aversion model, the comparative statics of the value of information are examined. This is an important check on the subjects' understanding of the fundamental exploration/exploitation tradeoff posed by bandit problems, and for verifying that the experimental design is powerful enough to test the hypotheses of interest.

4.1 Comparative statics

The most basic test of subjects' understanding is whether they assigned positive value to information they could acquire about the unknown arm. In the elicitation treatment, all 33 subjects expressed a positive willingness to pay for information about the unknown arm in all 6 bandits, even though the titration mechanism allowed them to express negative willingness to pay. Results from the Gittins index treatment were noisier. Table 2 shows the portion of the reported Gittins indexes that were greater than the expected value in each bandit, reflecting a positive value for information. In each bandit, between 20 and 23 of the 33 subjects selected a positive information value, rejecting the hypothesis that information values are strictly negative at $p \leq 0.08$ or better in four of the six bandits, and with $p < 10^{-4}$ in the sample overall. There is significant variation within subject, as only 12 subjects had positive information values in all six bandits, though 23 (sign test $p = 0.02$) had strictly positive median information values. Risk averse subjects could have Gittins indexes below expected value, but it is more likely that the complexity of the task introduced additional random variation. This complexity leads to positive information valuation rates that are

[12] Using $\lambda = 0.5$ in the willingness to pay treatment also maximizes the variance of the known arm, biasing risk averse subjects toward lower willingness to pay. Because the variance of Bernoulli distributions varies with θ, similar control is not possible in the Gittins index treatment, though in practice differences in variance are small. Importantly, it is the willingness to pay treatment where the unique prediction of ambiguity aversion is made, and this design provides the strongest control in that treatment, as any tendency to overpay is *despite* any risk aversion. The technique of controlling for risk aversion with lottery payments (Roth and Malouf 1979) relies on subjects' reducing compound lotteries, and is methodologically inconsistent when testing an ambiguity aversion model which suspends that same principle.

[13] The Caltech undergraduate population has a median SAT math score of 800, the maximum possible. Steyvers et al. (2009) associate IQ measures with higher rates of Bayesian strategies, suggesting this extraordinary mathematical and analytical ability presents a "best chance" test for the complex Gittins theory.

Table 2 Number of subjects with positive information values and associated p values of binary test that information values are positive

Unknown arm Prior (α, β)	A (1, 1)	B (2.5, 2.5)	C (1.1, 3.9)	D (3.9, 1.1)	E (2,3)	F (3,2)	Pooled
Positive	21/33	20/33	21/33	23/33	20/33	23/33	128/198
Zero	6/33	7/33	0	1/33	0	4/33	
p-value	0.08	0.15	0.08	0.02	0.15	0.02	10^{-4}

Table 3 Linear random effects of model of bandit attributes on observed Gittins indexes and Willingnesses to pay (t values in parentheses)

	Gittins index	Willingness to pay
Constant	−0.240	−0.151
	(−2.18)	(−0.73)
Prior mean	0.926	0.068
	(18.52)	(0.72)
Prior Std	0.395	1.114
	(1.61)	(2.42)
Log(periods)	0.019	0.187
	(0.53)	(2.79)
Wald χ^2(3 dof)	196.39	8.66

understandably slightly lower than the 80 to 90% observed in Banks et al. (1997) choice experiment.

In addition to expressing positive values for information, the observed magnitude of the value of information should increase with the horizon and the variance of the prior distribution. Table 3 uses a random effects model to estimate a linear approximation to how the two measures of the value of information vary with bandit characteristics. Theoretically, the Gittins index is the mean of the prior distribution, plus some additional value related to the variance and horizon of the bandit. Consistent with theory, the Gittins index reflects almost perfectly changes in the prior mean. Indexes also increase borderline significantly ($p = 0.108$) with increases in prior variance, reflecting the increased likelihood the unknown arm is much better than expected. While the coefficient on horizon is positive as predicted, it is not significant in this data ($p = 0.596$).

The observed willingnesses to pay are consistent with theory, as both the higher prior variance ($p = 0.016$) and longer horizons over which exploit information ($p = 0.005$) significantly increase subjects' willingness to pay. There is no predicted relationship between the prior mean and the willingness to pay, and none is identified in the model ($p = 0.474$).

While there is some noise in the observed Gittins indexes, the data suggest subjects recognize that information about the unknown arm has value, and that its value increases when there is a greater chance the unknown arm is much better than the known arm (higher variance) and when there are more opportunities to exploit the

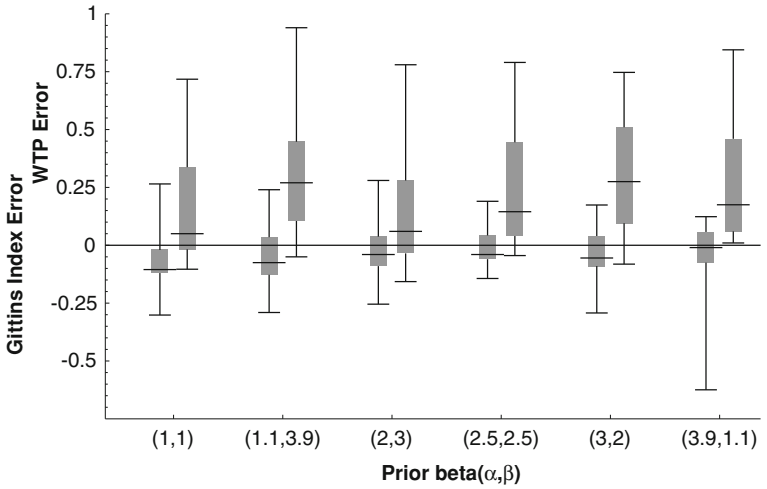

Fig. 1 *Box-and-whiskers* plot of the difference between subjects' responses and the optimal Gittins index (on the *left*) and the optimal willingness to pay (on the *right*)

better option (longer horizon). However, as the next section shows, the magnitude of the value placed on information is not optimal, and the deviations are consistent with the predictions of ambiguity aversion shown in Theorems 1 and 2.

4.2 Testing ambiguity aversion

Figure 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.[14]

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.

Table 4 presents the median overpayment and median undervaluation for each arm, as well as for the pooled data. Because ambiguity aversion predicts overpayment in the willingness to pay treatment and undervaluation in the Gittins index treatment, 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.

[14] Note that this is the distribution of the data, and does not correspond to confidence intervals of the central tendency of the data.

Table 4 One-tailed p-values that WTPs and Gittins indexes are optimal for each arm and for the pooled data, based on the median response

Unknown arm Prior (α, β)	A (1, 1)	B (2.5, 2.5)	C (1.1, 3.9)	D (3.9, 1.1)	E (2, 3)	F (3,2)	Pooled
Willingness to pay							
N	33	33	33	32	33	33	197
Optimal WTP	0.25	0.21	0.06	0.04	0.21	0.24	
Median overpay	0.05	0.15	0.29	0.19	0.08	0.29	0.17
# Overpay	24	29	30	29	23	31	166
p Value	0.0045	7×10^{-6}	1×10^{-6}	2×10^{-6}	0.012	2×10^{-7}	2×10^{-16}
Gittins index							
N	33	33	33	33	33	33	198
Optimal index	0.62	0.56	0.31	0.83	0.47	0.69	
Median undervalue	0.10	0.04	0.07	0.01	0.03	0.04	0.04
# Too low	27	19	24	19	20	22	131
p Value	0.0001	0.192	0.0045	0.192	0.112	0.0278	3×10^{-6}

For every bandit, the median overpayment is positive, meaning the median willingness to pay is higher than optimal. The fourth row indicates the number of overpayments which are positive, which can be used to calculate a p value for the hypothesis that the true median overpayment is equal to zero. The last row presents this p-value, which is significant at all conventional levels for each arm (p ranging from 0.012 to 2×10^{-7}), and extremely significant ($p < 2 \times 10^{-16}$) for the pooled data. Therefore, the median willingness to pay is significantly higher than optimal, consistent with the prediction of ambiguity aversion.

The lower section of the table presents the same information for undervaluation. The median undervaluation is positive in all six bandits, and significantly positive at conventional levels for three of the six bandits. However, the median of the pooled data is highly significantly positive ($p < 10^{-6}$), suggesting Gittins indexes are lower than optimal, consistent with the prediction of ambiguity aversion.

Taken together, the evidence in Table 4 offers strong support for ambiguity aversion playing a significant role in behavior within multi-armed bandit problems. Although the Gittins index elicitation procedure was complex and the data noisy, there is a consistent tendency toward undervaluing information that could be obtained through experimentation. However, the major novel prediction of ambiguity aversion, that willingness to pay to learn the true value of an unknown mean arm is higher than optimal, is very strongly supported.

In addition to these major predictions, ambiguity aversion makes subtle predictions about how the degree of overpayment or undervaluation will change with properties of the bandit. Looking at Fig. 1, the median level and range 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. Table 5 confirms

Table 5 Results of regressions of overpayment and undervaluation on prior mean and variance and the log of horizon

	Undervaluation (Gittins index)	Overpayment (Willingness to pay)
Constant	0.12	0.07
	(−0.85)	(0.31)
Prior mean	0.09	0.07
	(0.57)	(0.64)
Prior std	0.28	−0.52
	(0.84)	(−1.31)
Log(periods)	0.03	0.16
	(1.25)	(2.08)

t Statistics are in parentheses

this impression with a random effects regression of overpayment and undervaluation on the prior mean, prior standard deviation and log-horizon of the six bandits.

Within these data, the only significant effect is an increase in overpayment associated with longer horizons. While longer horizons provide more opportunities to exploit information acquired through purchase or experimentation, resulting in higher optimal Gittins indexes and willingnesses to pay, it is not obvious the degree to which ambiguity aversion predicts *over*payment and *under*valuation will vary with the horizon. However, this strong regularity suggests horizon variation may be a productive area for future theoretical analysis and testing.

The regressions suggest that there is no significant effect of the mean of the unknown mean arm, failing to offer support for a very subtle prediction of ambiguity aversion. Although the degree of undervaluation should not be affected by the prior mean, ambiguity aversion predicts willingness to pay should be exactly the same as that of ambiguity neutral agents when the mean is below 0.5: when the Gittins index is below 0.5, both the ambiguity averse and ambiguity neutral agents expect to pick the 50/50 known arm in the first period, so the bandit involves no ambiguity, and the value of knowing the mean of the ambiguous arm is the same. Therefore, the level of overpayment may increase at higher means, over 0.5, since those players will actually face the ambiguous alternative.

The one intuitively obvious effect of ambiguity aversion is that an increase in prior variance—an increase in ambiguity—should increase overpayment and undervaluation. Within these data, there is no systematic effect of the variance of the prior. While undervaluation has the correct sign, the sign on overpayment actually suggests people are willing to pay (insignificantly) more when there is less ambiguity. In comparing the (1, 1) bandit (prior standard deviation 0.289) with the (2.5, 2.5) bandit (prior standard deviation 0.204), undervaluation is higher for 25 of 33 subjects in the more ambiguous bandit ($p = 0.003$), but overpayment is actually lower for 20 subjects ($p = 0.15$). The relatively small difference in standard deviation may simply be swamped by other sources of error and variation in the data.

5 Discussion

How people approach multi-armed bandit problems is key to understanding behavior in circumstances of sequential choice under uncertainty, such as brand choice,

natural resource exploration, research and development, and job and price search. While previous choice-based experiments suggest people do appreciate that there is value to information obtained through experimenting with less familiar alternatives, choice patterns are not optimal in the sense that backward-looking switching models are more predictive than Gittins theory. Further evidence from field and laboratory search problems suggests people do undervalue the possibility of further exploration, preferring to accept the proverbial bird in the hand. In this article, it is shown that these empirical regularities can be explained by ambiguity aversion, a model that also makes a striking prediction that people will be willing to pay more than optimal to learn the true mean of ambiguous alternatives. Since most other explanations for, or ad hoc characterizations of, bandit behavior do not make this prediction, it provides a particularly strong test of whether ambiguity aversion plays a role in the multi-armed bandit environment.

Results of an experiment designed to elicit Gittins indexes and willingnesses to pay for information in two-armed bandits offer strong support for ambiguity aversion playing a dominant role in bandit decision making. Data in the Gittins index treatment are broadly consistent with previous results that people do recognize that information gained through experimentation can improve their future payoffs. In addition, Gittins indexes are overwhelmingly smaller than optimal, consistent with ambiguity aversion. Observations in the willingness to pay treatment indicate that willingness to pay is positive in every observation, and consistent with ambiguity aversion, higher than optimal in an overwhelming portion of the data. Models which consider agents as risk averse (Banks et al. 1997) or considering only a shortened-horizon version of the problem (e.g., Meyer and Shi 1995; Acuna and Schrater 2008) correspondingly discount the value of learning the true mean of the unknown alternative, and assign a lower willingness to pay. Taken together, these results present compelling support for ambiguity aversion, to the exclusion of most other models which have been considered in understanding bandit behavior.

While there is noise in the data, likely arising from the complexity of the elicitation tasks, it does little to weaken support for the ambiguity aversion model. Most of the noise is in the Gittins index data, and it is in this domain where previous research leaves little doubt that people underexperiment. The compelling argument for ambiguity aversion arises in the willingness to pay treatment, where overpayment is not predicted by most other theories, and data patterns from this experiment are quite clear.

Awareness of agents' aversion to ambiguity suggests opportunities for market and policy responses that improve outcomes in naturally occurring bandit problems. In brand choice problems, companies introducing new brands might encourage experimentation by offering free samples at the supermarket or through the mail, or generous coupons. That consumers are also willing to pay for information creates a market for information which reduces ambiguity. *Consumer Reports* and J.D. Power & Associates publish reliability information for major purchases which reduce ambiguity, rather than risk, because they use a large sample to provide a good estimate of how likely a particular model is to be troublesome, which does not affect the model-conditional probability of getting a defective product.

While the market can respond productively to ambiguity aversion in many settings, that ambiguity aversion affects bandit behavior also has implications for public policy. For instance, tying the size and duration of unemployment insurance benefits to the job market success of other job seekers helps insure against bad average job market outcomes and encourages ambiguity averse job seekers to continue searching, leading to lower levels of underemployment. Ambiguity aversion also creates a policy need to ensure underexperimenting agents are not exploited, such as by bait-and-switch scams, or misleading advertising.

Addressing the effects of ambiguity aversion through policy, or allowing the market to do so without limitation, runs the risk that the welfare loss caused by any resulting overexperimentation exceeds the loss caused by initial underexperimentation. However, an environment that leads to overexperimentation, or overpayment for information rather than underpayment, is preferable to one which abides underexperimentation for two reasons. First, in many contexts, such as price search, experimentation is a public good; while overexperimentation is privately costly, positive externalities, like very competitive markets, offset some of this cost. Second, overexperimenting or overpaying agents have both the incentive and the *information* to correct their behavior. Once she has bought the new orange juice a second time, a consumer has the opportunity to regret her purchase and modify her behavior. It is difficult to learn to experiment more, however, because the underexperimenting agent does not have the information necessary to determine that she is not optimizing; she does not know what she is missing.

This work has suggested several avenues that may provide additional insight into behavior in the bandit environment. One open question is how framing the bandit as a value elicitation problem, rather than a choice problem, affects behavior. Continuing work suggests the value elicitation results generalize, but that the mechanism used to compute indexes is not the same used to solve choice-based bandit problems. Further inquiry into index strategy computation could examine simple environments, where computing an index is a relatively easy task, to ensure the results presented here are not an artifact of some computational shortcut. Another issue places the bandit problem in a market context: how much search needs to occur for competitive, or near competitive market outcomes to obtain? and can price convey enough information that agents are guided to make nearly optimal choices (i.e., is expected surplus equal across alternatives)? These questions could be addressed in the laboratory, or using field data such as supermarket scanner panels.

In many economic problems, optimal solutions are too complex to calculate to realistically believe that agents compute optimal solutions. In such cases, it is hoped that agents can approximate solutions well enough to nevertheless achieve nearly optimal outcomes. In bandits, although observed strategies reflect many comparative statics of optimal behavior, the magnitudes of the observed indexes suggest behavior systematically differs from optimality. In these cases, policies which will improve agents' welfare must address the cause of the suboptimality. Understanding these causes requires developing and testing a variety of sensible alternative models which may explain the observed patterns. Based on the theoretical and experimental evidence in this paper, ambiguity aversion is consistent with what is known about bandit behavior, but also accurately predicts behavior in the untested domain of willingness to pay

for information. This strong support suggests it may be worthwhile to identify cases where ambiguity may be acting in naturally occurring bandits and seek to implement policies which minimize its welfare effects.

Acknowledgments I am grateful for funding from the Russell Sage Foundation, and to Jeff Banks, Colin Camerer, Rachel Croson, Paolo Ghirardato, and several anonymous reviewers for helpful comments.

References

Acuna, D., & Schrater, P. (2008). Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In V. Sloutsky, B. Love, K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society*. Washington, DC: Cognitive Science Society.

Anderson, C. (2001). Behavioral models of strategies in multiarmed bandit problems. California Institute of Technology Ph.D. Dissertation.

Anscombe, F., & Aumann, R. (1963). A definition of subjective probability. *Annals of Mathematical Statistics, 34*, 199–205.

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

Berry, D., & Fristedt, B. (1985). *Bandit problems*. New York: Chapman and Hall.

Boyce, J., Bruner, D., & McKee, M. (2009). *Be my guinea pig: Information spillovers in a one-arm bandit problem*. Appalachian State University Department of Economics working paper.

Braunstein, Y., & Schotter, A. (1982). Labor market search: An experimental study. *Economic Inquiry, 20*, 134–144.

Camerer, C. (1995). Individual decision making. In A. Roth & J. Kagel (Eds.), *Handbook of experimental economics*. Princeton: Princeton University Press.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk AndUncertainty, 5*, 325–370.

Cox, J., & Oaxaca, R. (1989). Laboratory experiments with a finite-horizon job-search model. *Journal of Risk and Uncertainty, 2*, 301–329.

Cox, J., & Oaxaca, R. (1990). Unemployment insurance and job search. In Research in labor economics (Vol. 11, pp. 223–240). Greenwich, CT: JAI Press.

Cox, J., & Oaxaca, R. (1992). Direct tests of the reservation wage property. *The Economic Journal, 102*, 1423–1432.

Cox, J., & Oaxaca, R. (1996). Testing job search models: The laboratory approach. In S. W. Polachek (Ed.), Research in labor economics (Vol. 15, pp. 171–207). Greenwich, CT: JAI Press

Cox, J., & Oaxaca, R. (2000). Good news and bad news: Search from unknown wage offer distributions. *Experimental Economics, 2*, 197–226.

de Finetti, B. (1977). Probabilities of Probabilities. In A. Aykac & C. Brumat (Ed.), New directions in the application of Bayesian methods (pp. 1–10). Amsterdam: North Holland.

Einhorn, H., & Hogarth, R. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychology Review, 92*, 433–461.

Ellsberg, D. (1961). Risk ambiguity and the savage axioms. *Quarterly Journal of Economics, 75*, 643–669.

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1*, 149–157.

Gans, N., Knox, G., & Croson, R. (2007). Simple models of discrete choice and their performance in bandit experiments. *Manufacturing and Service Operations Management, 9*, 383–408.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics, 18*, 141–153.

Gittins, J. (1989). *Multi-arm bandit allocation indicies*. New York: Wiley.

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani et al. (Ed.), *Progress in statistics* (pp. 241–266). Amsterdam: North Holland.

Halevy, Y. (2007). Ellsberg revisited: An experimental study. *Econometrica, 75*, 503–536.

Hey, J. (1987). Still searching. *Journal of Economic Behaviorand Organization, 8*, 137–144.

Hogarth, R., & Einhorn, H. (1990). Venture theory: A model of decision weights. *Management Science, 36*, 780–803.

Kahn, B., & Sarin, R. (1988). Modeling ambiguity in decisions under uncertainty. *Journal of Consumer Research, 15*, 265–272.

Klabinoff, P., Marinacci, S., & Mukerji, M. (2009). Recurseive smooth ambiguity preferences. *Journal of Economic Theory, 144*, 930–976.

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

Norton, D., & Isaac R. (2010). *Experts with a conflict of interest: A source of ambiguity?*. University of Florida, Department of Economics working paper.

Pratt, J., Wise, D., & Zeckhauser, R. (1979). Price differences in almost competitive markets. *Quarterly Journal of Economics, 94*, 189–211.

Roth, A., & Malouf, M. (1979). Game theoretic models and the role ofinformation in bargaining: An experimental study. *Psychological Review, 86*, 574–594.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571–587.

Schotter, A., & Braunstein, Y. (1981). Economic search: An experimental study. *Economic Inquiry, 19*, 1–25.

Segal, U. (1987). The Ellsberg paradox and risk aversion: An anticipated utility approach. *International Economic Review, 28*, 175–202.

Steyvers, M., Lee, M., & Wagenmakers, E.-L. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53*, 168–179.

Traeger, C. (2010). *Subjective risk, confidence and ambiguity*. University of California, Berekely Department of Agricultural and Resource Economics Working Paper 1103.

Yates, F., & Zukowski, L. (1976). Characterization of ambiguity in decision making. *Behavioral Science, 21*, 19–25.