Experimental Economics (2025), 1–23
doi:10.1017/eec.2025.10027

ORIGINAL PAPER

Strategies in the multi-armed bandit

Stanton Hudja[1] and Daniel Woods[2]

[1]Stuart School of Business, Illinois Institute of Technology, Chicago, Illinois, USA
[2]Department of Economics, MQBS Experimental Economics Laboratory, Macquarie Business School, Sydney, New South Wales, Australia
Corresponding author: Daniel Woods; Email: daniel.woods@mq.edu.au

(Received 6 January 2023; revised 3 September 2024; accepted 29 July 2025)

Abstract
This paper analyzes individual behavior in multi-armed bandit problems. We use a between-subjects experiment to implement four bandit problems that vary based on the horizon (indefinite or finite) and the number of bandit arms (two or three). We analyze commonly suggested strategies and find that an overwhelming majority of subjects are best fit by either a probabilistic "win-stay lose-shift" strategy or reinforcement learning. However, we show that subjects violate the assumptions of the probabilistic win-stay lose-shift strategy as switching depends on more than the previous outcome. We design two new "biased" strategies that adapt either reinforcement learning or myopic quantal response by incorporating a bias toward choosing the previous arm. We find that a majority of subjects are best fit by one of these two strategies but also find heterogeneity in subjects' best-fitting strategies. We show that the performance of our biased strategies is robust to adapting popular strategies from other literatures (e.g., EWA and I-SAW) and using different selection criteria. Additionally, we find that our biased strategies best fit a majority of subjects when analyzing a new treatment with a new set of subjects.

Keywords: Experimentation; Multi-armed bandits; Strategy selection; Reinforcement learning

JEL Codes: C91; D83; O30

> "Bandit problems embody in essential form a conflict evident in all human action: information versus immediate payoff." – P. Whittle in Gittins (1989)

1. Introduction

The tension between immediate reward maximization and information is universal in decision-making. For example, a consumer often chooses between their favorite brand and a new brand that may reveal itself to be better. Similarly, a researcher often chooses between their current research agenda and a potentially promising new area of research. In each of these examples, an individual has to weigh the short-term benefit of immediate reward maximization (exploitation) versus the long-term benefit of information (exploration). Other examples of this exploration versus exploitation trade-off occur in many other environments, such as resource exploration, experimental design, and job search.

Despite the prevalence of this exploration versus exploitation trade-off, it is unclear how individuals actually resolve this trade-off when presented with it. In this paper, we use a series of experiments

to uncover how individuals resolve this trade-off. We do this by using the canonical multi-armed bandit problem (Thompson, 1933) to represent the exploration versus exploitation trade-offs found in everyday decision making.[1] The multi-armed bandit problem is analogous to a gambler continually choosing between various slot machines (or one-armed bandits) that each have an unknown reward distribution. The exploration versus exploitation trade-off arises as the gambler, given her information, must continually decide between the machine with the highest immediate expected reward and learning about the other machines.

In this paper, we explore the strategies that individuals use in multi-armed bandit problems. We ask two main questions in this study. First, which of the previously suggested strategies in the bandit literature best capture subject behavior? We compare previously suggested strategies from the computer science and management literatures to see how many subjects they best describe. Second, can we improve upon these previously suggested strategies using the data from the experiment? We build upon the best-fitting strategies by adapting them to fit important behavioral trends that these strategies are not capturing. It is important to address these questions as uncovering the strategies that best describe subject behavior will help us better understand how individuals resolve experimentation problems.

We design an experiment to analyze subject behavior in multi-armed bandit problems. In our baseline treatment, the two-armed indefinite horizon treatment, subjects repeatedly draw a ticket from one of two boxes. Each box has a constant probability of returning a ticket that pays out a constant reward. Subjects know that each box's payout probability is drawn from a standard uniform distribution but do not know the payout probability for either box. Subjects repeatedly draw a ticket from either one of the boxes until the random termination of the bandit problem.

An individual in this problem repeatedly faces an exploration versus exploitation trade-off. An individual should initially believe that each box is equally likely to pay out a reward, as each box has a reward rate drawn from the same distribution. However, as an individual samples from a box, their belief of that box's reward rate should change as they observe the outcomes from drawing from that box. As an individual's beliefs change, they must frequently decide between drawing from the box that has the highest expected reward rate and learning more about the other box. An individual is maximizing their expected immediate reward by choosing the box with the highest expected reward rate. However, an individual may be better off in the future learning about the other box as this box has some possibility of having a higher true reward rate. The focus of our study is to determine how a subject chooses between these options.

The experiment has three other treatments besides the baseline treatment: (i) the two-armed finite horizon treatment, (ii) the three-armed indefinite horizon treatment, and (iii) the three-armed finite horizon treatment. In the finite horizon treatments, subjects have a fixed known number of tickets that they can draw. In the three-armed treatments, subjects can draw tickets from a third box. These treatments allow us to uncover patterns of behavior that are consistent across various bandit problems and uncover how robust strategies are to changes in the environment. We vary the type of time horizon as behavior may depend on whether the expected number of future decisions is held constant. We vary the number of bandit arms as implementing certain strategies may be more cognitively taxing in three-armed bandit problems, given the increased number of options.

We first analyze subject behavior to classify trends that will be compared to our strategies. We classify subject behavior based on how often subjects choose a myopically suboptimal action (experiment), switch between actions, and choose the best box (i.e., the box with the highest true underlying reward rate). After controlling for the number of successes and failures from each box, we find that (i) subjects are more likely to choose a given box when it was chosen last and (ii) subjects are less likely to switch following a success than a failure. These results suggest that, after controlling for the information from each box, subjects treat arms differently based on whether they were last chosen and

[1]The specific version analyzed in this paper was established in Robbins (1952).

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

treat a previous success differently than a previous failure. We also find that subjects are less willing to switch over time conditional on the last outcome. This suggests that subjects make choices based on more than the previous outcome.

We then compare the fit of previously suggested strategies on subject behavior. We consider 14 strategies that have either been suggested in the computer science literature or the operations management literature. We fit each strategy for each subject using Maximum Likelihood Estimation, and we compare each strategy for each subject using the Bayesian Information Criterion (BIC), which penalizes the log-likelihood for each strategy based on the number of free parameters.[2] Under this penalization, when comparing two strategies, the strategy with more parameters must have a sufficiently larger log-likelihood for it to be selected by BIC.

We find that an overwhelming majority of subjects in the experiment are classified as using either a probabilistic win-stay lose-shift strategy or a reinforcement learning strategy when we compare previously suggested strategies. In the probabilistic win-stay lose-shift strategy, the probability that a subject switches from the previous arm depends solely on the previous outcome; a subject can switch with a different probability following a success than following a failure. In reinforcement learning, subjects are more likely to play arms that return successes relatively more often. While these strategies fit best among previously suggested strategies, there is evidence that subjects are not playing these strategies. For example, we show that subjects classified as using probabilistic win-stay lose-shift are less likely to switch over time (conditional on the last outcome). Thus, these subjects appear to violate the assumptions of the probabilistic win-stay lose-shift strategy as their switching behavior depends on more than just the last outcome.

We propose three new strategies that build on previous strategies by incorporating behavioral patterns from the data. As we observe decreasing switching rates over time, we first estimate a decreasing win-stay lose-shift strategy where subjects are less willing to switch, conditional on the previous outcome, over time. The second and third new strategies are "biased" strategies that incorporate a bias toward the previous choice. The second new strategy is biased reinforcement learning, where subjects are reinforcement learners who are biased towards their previous choice. The third new strategy is the biased myopic strategy, where subjects are generally myopic, but are biased toward their previous choice. The biased strategies place arm evaluations (i.e., reward-based indices used in choice) and inertia into a logit function. Biased myopic uses expected reward to evaluate arms, while biased reinforcement learning uses propensity. An overwhelming majority of subjects in each treatment are best fit by one of these two biased strategies.

We run various robustness checks on our results. First, we adapt the Inertia Sampling and Weighting (I-SAW) model (based on Nevo & Erev (2012)), the Experience-Weighted Attraction (EWA) model (based on Camerer & Ho (1999)), and the self-tuning EWA model (based on Ho et al. (2007)) to our environment and estimate them. We still find that our biased strategies best fit a large majority of subjects. Second, we show that our results are qualitatively unchanged by using a selection criterion that is not based on log-likelihood. Lastly, we find that our biased strategies best fit a majority of subjects in a new treatment that consists of a new set of subjects. The consistent performance of the biased strategies in our paper makes two important suggestions. First, it suggests that subjects' behavior depends on how similarly they evaluate the arms to be. As subjects appear to evaluate arms using a reward-based process, we expect decreased experimentation when arms have returned rewards at very different rates. Second, it suggests that while subjects tend to incorporate all observed outcomes into their decision-making, they treat arms differently based on whether they were last chosen. Our biased strategies differ from the rest of the strategies that we estimate by directly modeling both of these suggestions.

[2]Our results are qualitatively similar using alternative selection criteria such as log-likelihood, the Akaike Information Criterion (AIC), and the Brier Score. We report this analysis in Online Appendix J.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

This paper contributes to three main strands of literature. The first is the literature on multi-armed bandit experiments.[3,4] Previous experiments have generally made design choices that complicate analysis of the classic multi-armed bandit problem. Horowitz (1973) analyzes a two-armed finite horizon bandit problem but does not inform subjects how the bandit arms are generated. Banks et al. (1997) deviate from the classic multi-armed bandit problem by analyzing a special case of the two-armed indefinite horizon bandit problem where subjects are always predicted to be myopic. Anderson (2001) analyzes a four-armed bandit problem where payoffs are drawn from a normal distribution but draws payoffs from a different distribution than the one that was implied to subjects. Lastly, Gans et al. (2007) analyze the selection of some deterministic strategies in a two-armed bandit problem but provide subjects with inaccurate initial information on each bandit arm.[5] They find that hot-hand strategies fit best out of the deterministic strategies they estimate. Gans et al. (2007) is the paper closest to our research. We differ from this paper by analyzing the multi-armed bandit problem without deception, by analyzing multiple multi-armed bandit environments, by incorporating probabilistic strategies, and by building new strategies based on the data moments. We find that hot-hand strategies no longer fit best once probabilistic strategies are included.

Our paper contributes to the literature on multi-armed bandit experiments in a few ways. First, we develop new strategies that better capture individual subject behavior than commonly suggested bandit strategies and strategies that we adapt from other literatures. Our biased strategies differ from other strategies that we estimate by directly modeling both (i) probabilistic choice that depends on the relative differences in arm evaluations and (ii) inertia.[6] Our biased strategies parsimoniously incorporate these two features by placing arm evaluations (either expected reward or propensity) and an inertia constant(s) into a logit function. Second, we show that the best-fitting strategies appear to be similar under different types of horizons and when comparing two to three arms. Thus, subjects' underlying behavior may be robust to modest changes in the environment. Lastly, we reintroduce the canonical multi-armed bandit problem as a tool that experimental economists can use to analyze the exploration versus exploitation trade-off. There is a scarcity of experimental economics research on this problem, considering it presents a trade-off that is inherent in many environments. One previous barrier to conducting these studies was the difficulty in obtaining optimal predictions. We demonstrate how approximation methods, coupled with simulations, can provide these predictions.

The second strand of literature is on reinforcement learning in economic experiments. Reinforcement learning became popular in economics following the seminal work of Roth & Erev (1995), who show that reinforcement learning can explain differences in games with similar equilibria. Since Roth & Erev (1995), there have been papers that continue to analyze how well reinforcement learning can capture behavior in various games (Erev & Roth, 1998; Feltovich, 2000) and papers that have augmented reinforcement learning with plausible behavioral considerations

[3]There is a larger literature on one-armed bandit experiments. Some of these papers are Anderson (2012), Banks et al. (1997), Banovetz & Oprea (2023), Deck & Kimbrough (2017), Hoelzemann & Klein (2021), Hudja (2019), Hudja & Woods (2024), Hudja (2021), and Meyer & Shi (1995). In these bandit problems, only one arm has an unknown reward distribution. These problems tend to theoretically reduce to stopping problems, which differ from the problems we analyze in our study.

[4]Other bandit problems have been experimentally analyzed in different disciplines. These bandits either do not have theoretical predictions or differ from the exploration versus exploitation trade-off found in our bandit problems. A few papers analyze (restless) bandits that have arms with nonstationary reward distributions (Daw et al., 2006; Yi et al., 2009; Addicott et al., 2013; Speekenbrink & Konstantinidis, 2015; Navarro et al., 2018; Hotaling et al., 2021). Some papers analyze bandits where the arms are correlated (Gershman & Niv, 2015; Wu et al., 2017, 2018; Schulz et al., 2020). Other bandits can be found in Toyokowa et al. (2014), von Helversen et al. (2018), Kip Viscusi & DeAngelis (2018), and Gershman (2019).

[5]Gans et al. (2007), in each bandit problem, tell subjects that each of the two arms in the bandit problem has been sampled three times and that each arm has had two successes out of those three trials. This type of inaccurate information may result in subjects doubting the veracity of the instructions. We avoid this design choice by providing accurate information in our instructions. In general, misrepresenting or leaving out information on the prior complicates testing subject behavior.

[6]One strategy that we do not estimate is from Ferecatu & De Bruyn (2022). This model can capture similar features but would be computationally impractical using our estimation approach.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

(Duffy & Feltovich, 1999; Rosokha & Younge, 2020). Rosokha & Younge (2020) is a notable example as they incorporate loss aversion into reinforcement learning to show that loss aversion can lead to greater persistence at exploration (in a non-bandit environment). One area where reinforcement learning has become popular is in decisions from experience (as defined by Erev & Haruvy (2016)). In these decisions from experience, individuals must choose between options when they are not given any prior description of the incentive structure. Decisions from experience differ from multi-armed bandit problems in that decisions from experience do not have an explicit exploration versus exploitation trade-off. In the complete absence of description, expected payoffs cannot be calculated or inferred; subjects cannot even be sure that the payoff distribution is stationary. Nevo & Erev (2012) propose a type of reinforcement learning model ("I-SAW") that appears to capture behavior well in these environments. Our biased strategies share some similarities with I-SAW as they both allow for subjects to treat previously chosen arms differently from previously unchosen arms.[7] However, I-SAW is not directly applicable to our environment without making substantive changes to account for the lack of forgone payoffs. After making these changes, we find that our biased strategies better fit more subjects than a modified I-SAW.[8] Our paper contributes to the literature of reinforcement learning in economics experiments by showing that reinforcement learning models best fit the most subjects (out of the models we consider) in multi-armed bandit environments. Additionally, we adapt reinforcement learning by incorporating a bias toward the previous choice and show that this model best fits a plurality of subjects.

The third strand of literature is the literature on strategy selection in economics experiments. There are many papers on strategy selection that focus on behavior in the indefinitely repeated prisoner's dilemma game. Dal Bó & Fréchette (2011) use a finite mixture model to show that subjects, depending on the parameters, tend to use the always defect or tit-for-tat strategies. Various subsequent papers have analyzed strategies in this environment (Fudenberg et al., 2012; Romero & Rosokha, 2018; Romero & Rosokha, 2023). These papers have shown that the tit-for-tat, grim-trigger, and always defect strategies are quite common. Our paper contributes to this literature by analyzing strategies in the multi-armed bandit problem, which has a unique short-term versus long-term trade-off of immediate reward maximization versus information.

2. Multi-armed bandit problem

This section consists of three subsections. The first subsection describes the specific bandit problems we consider. The second subsection explains optimal behavior in our indefinite horizon bandit problems. The third subsection explains optimal behavior in our finite horizon bandit problems.

2.1. Multi-armed bandit problems

The multi-armed bandit problem was first proposed by Thompson (1933, 1935) as a problem of determining which of two medical treatments is superior (in a timely manner). The multi-armed bandit problem consists of N arms that can be pulled in any order and at any time. Each pull from an arm results in a reward randomly drawn from a fixed distribution. An agent's objective in the multi-armed bandit problem is to maximize their expected value of rewards given the time horizon.

[7] There are substantive differences between I-SAW and our models. I-SAW differs from our biased models by modeling a random behavior phase where choices are randomly made, independent of relative arm evaluations. I-SAW has inertia that directly depends on the surprise from recent outcomes, while our biased models do not use a surprise factor. Additionally, I-SAW has an implicit underweighting of rare events due to its reliance on small samples; this underweighting is not captured in our biased models as it is not as appropriate for our environment.

[8] In subsection 6.4, we modify I-SAW for our environment and show that our biased strategies fit better than this modified I-SAW. In this modified strategy, we adjust I-SAW's "surprise", "inertia", and "sampling" definitions to reflect the fact that in multi-armed bandit subjects only observe the outcome of the arm that was chosen. We also further adapt I-SAW as I-SAW's original simulation-based approach is not computationally feasible using the estimation approach used in this paper.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

We study a common version (Robbins, 1952) of the multi-armed bandit problem where each arm is a Bernoulli process with a different unknown success probability. An arm returns a value of 1 in the event of a success and a value of 0 in the event of a failure. The constant probability (θ_i) of arm i returning a success is drawn from a beta distribution with parameters α_0 and β_0. After each draw from an arm, an agent should update her beliefs about that arm through Bayesian updating. After s_i successes and f_i failures on arm i, the expected reward for arm i, given the beta distribution, is $\frac{s_i+\alpha_0}{s_i+f_i+\alpha_0+\beta_0}$.

In the experiment, there are four specific bandit environments. The first environment is the two-armed indefinite horizon bandit problem. In this environment, a subject faces two unknown arms that each has a probability of success drawn from a beta distribution with $\alpha_0 = 1$ and $\beta_0 = 1$.[9] An agent in this environment discounts future rewards by a discount factor of $\delta = 0.96$. The second environment is the two-armed finite horizon bandit problem, which is similar to the previous environment except that a subject can only make 25 decisions (the same expected number of decisions as the previous environment). The third and fourth environments are three-armed versions of the two-armed indefinite horizon bandit problem and the two-armed finite horizon bandit problem.

2.2. Optimal behavior for indefinite horizon

Gittins & Jones (1974) and Gittins (1979) show that an optimal solution to the indefinite horizon multi-armed bandit problem is to always choose the arm with the largest Gittins index. The Gittins index for an arm is the maximum discounted expected reward per unit of discounted time. Let $r(X_t)$ denote the reward associated with the observation X_t and π be the state of the arm. For each arm i, the following index is calculated,

$$\nu(\pi) = \sup_{\tau} \left\{ \frac{E \sum_{t=0}^{\tau-1} \delta^t r(X_t)}{E \sum_{t=0}^{\tau-1} \delta^t} \right\},$$

where τ is a stopping time. The idea is to find for each arm the stopping time τ that results in the highest discounted expected return per discounted expected number of rounds in operation. After finding this τ for each arm, the Gittins index for each arm is compared, and the arm with the largest Gittins index is chosen.

While Gittins indices solve the indefinite horizon bandit problem, the state space is too large to compute exact Gittins indices. Thus, we use an approach suggested by Wang (1997) that allows us to approximate Gittins indices. We outline this approach in Online Appendix E. Wang (1997) provides bounds on the approximation error for using this approach. Through this approximation, we are able to calculate "approximate Gittins indices" that are within $5e^{-8}$ of the actual Gittins indices. This allows us to form predictions through simulation by using approximated Gittins indices.

2.3. Optimal behavior for finite horizon

The solution for a finite horizon bandit problem does not rely on Gittins indices. The solution can be derived using dynamic programming. We use the two-armed finite horizon problem as an example. Denote the value function in the last decision (T) as $V(T, s_1, f_1, s_2, f_2)$. In the last decision, the expected value from drawing from arm i is given by $\frac{s_i+\alpha_0}{s_i+f_i+\alpha_0+\beta_0}$. Thus, it is optimal to choose the arm with the highest expected payoff. The value function in decision T is given by

$$V(T, s_1, f_1, s_2, f_2) = max \left\{ \frac{s_1 + \alpha_0}{s_1 + f_1 + \alpha_0 + \beta_0}, \frac{s_2 + \alpha_0}{s_2 + f_2 + \alpha_0 + \beta_0} \right\}.$$

[9]This distribution is also known as the standard uniform distribution.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

In a decision (t) before T, an agent should consider the possibility of learning about the arm that they choose. Let $\mu_1 = \frac{s_1+\alpha_0}{s_1+f_1+\alpha_0+\beta_0}$ and $\mu_2 = \frac{s_2+\alpha_0}{s_2+f_2+\alpha_0+\beta_0}$. The value function in decision t is given by

$$V(t, s_1, f_1, s_2, f_2) =$$
$$max\{\mu_1(1 + V(t+1, s_1+1, f_1, s_2, f_2)) + (1-\mu_1)V(t+1, s_1, f_1+1, s_2, f_2),$$
$$\mu_2(1 + V(t+1, s_1, f_1, s_2+1, f_2)) + (1-\mu_2)V(t+1, s_1, f_1, s_2, f_2+1)\}.$$

We can use these value functions and backward induction to calculate the optimal decision for each possible value of t, s_1, f_1, s_2, and f_2. We will use these optimal decisions in simulations to make predictions.

3. Experimental design

The experiment is designed with two goals in mind. The first goal is to uncover the strategies that subjects use in multi-armed bandit problems. The second goal is to uncover whether strategy use is robust to changes in the environment.

3.1. Treatments and parameters

The experiment consists of four treatments: (i) the two-armed indefinite horizon treatment, (ii) the two-armed finite horizon treatment, (iii) the three-armed indefinite horizon treatment, and (iv) the three-armed finite horizon treatment. We use a between-subjects design where each subject faces 30 periods (i.e., 30 bandit problems) in the same environment. In the indefinite horizon treatments, the discount factor (continuation probability) is $\delta = 0.96$, which results in an expected period length of 25 rounds (arm pulls). In the finite horizon treatments, the period length is 25 rounds. In each treatment, the payoff from a success is \$0.50 and the payoff from a failure is \$0.00.[10]

3.2. Experiment

Instructions for the experiment were displayed on each subject's computer. After subjects read the instructions, they completed five incentivized comprehension questions. Upon completion of the comprehension questions, the experiment began.

We use the two-armed indefinite horizon treatment as an example of the experiment. In the two-armed indefinite horizon treatment, subjects must repeatedly draw a ticket from either the "L box" or the "R box". Each box has 1,000 tickets, which are a random combination of red and blue tickets. At the start of the period, the number of red tickets in each box is randomly chosen, with each integer between 0 and 1,000 being equally likely.[11] If a subject draws a red ticket from a box, they obtain \$0.50. If a subject draws a blue ticket from a box, they obtain \$0.00. Tickets are drawn with replacement, so that once a ticket is drawn from a box, it is placed back in the appropriate box. A subject repeatedly chooses between the L box and the R box until the bandit problem randomly ends due to the random termination probability.[12]

Figure 1 shows an example of the interface for the two-armed indefinite horizon treatment. For each box, the interface displays the number of red tickets drawn so far, the number of blue tickets drawn so far, and the last ticket drawn. In addition to this information, the interface has history buttons for each box that display a table with the previous outcomes of that box when pressed. These history boxes can update on their own when opened, and it was possible for subjects to have all

[10] The scaling of payoffs in this way does not affect predictions using the optimal strategy.

[11] These probabilities are limited to the third decimal place for ease of explanation while satisfying the constraint of the discretization having a negligible effect on our predictions (see Online Appendix F).

[12] All random variables were generated during the experiment. This is important so that our strategy estimation is not influenced by a specific realization of random variables that everyone faces.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

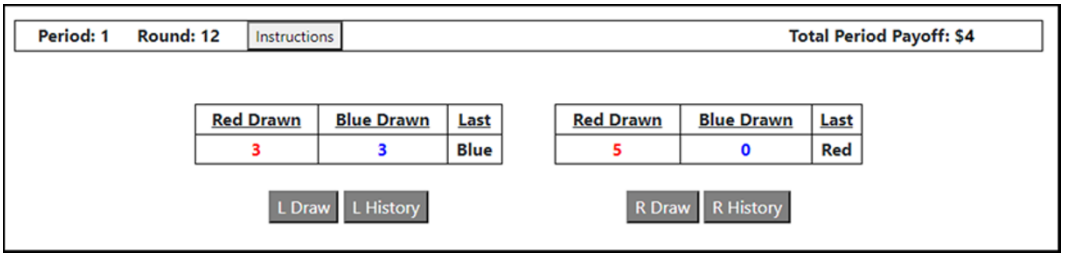

Fig. 1 An example of the experimental interface for the two-armed indefinite horizon treatment

of these history boxes open when making decisions. The interface also displays the period (bandit problem), the round (the current draw number), a button that displays a recap of the instructions, and the total period payoff.

The other treatments differ slightly from the two-armed indefinite horizon treatment. The two-armed finite horizon treatment has a known ending of 25 rounds for each period. The three-armed indefinite horizon treatment and the three-armed finite horizon treatment are similar to their two-armed counterparts, except that there is a third box (the "M box") that subjects can draw from that is generated in the same way as the other boxes.

3.3. Procedures

The experimental sessions were conducted on Prolific (prolific.co).[13] The data was collected in June 2021. We recruited subjects who were between 18 and 27 years of age, who lived in the United States, and who were in college. We wanted to recruit subjects with similar demographics to subjects who would be found on a college campus so that our sample would be similar to typical physical laboratory-based studies. There are 215 subjects in the experiment, with 53 subjects in the two-armed finite horizon treatment and 54 subjects in the other treatments. Subjects were paid a completion fee of $3.50, paid $0.20 for each correct answer to the five comprehension questions, and paid for one random period (bandit problem) of the experiment. Subjects earned $12.33 on average for an experiment that generally lasted between 15 and 25 minutes.[14]

4. Predictions

In this section, we present predictions for each treatment. These predictions were obtained through simulation. For each treatment, each simulation consisted of one million bandit problems based on optimal behavior. In the indefinite horizon treatments, we use approximated Gittins indices to simulate behavior.[15] In the finite horizon treatments, we solved for the optimal decision in each possible situation through backward induction and used these optimal decisions in the simulation. Details on the simulation can be found in Online Appendix E. In this section, we report the averages of these simulated bandit problems.[16]

[13] Subjects in online experiments have been shown to behave similarly to subjects in laboratory experiments (Arechar et al., 2018; Snowberg & Yariv, 2021). The experiment was coded in oTree (Chen et al., 2016).

[14] The effective hourly rate is clearly highly incentivized for Prolific, who recommended paying at least $9.60 an hour at the time that the experiment was run.

[15] Our simulations are unaffected by the approximation. There is no decision where the approximated Gittins indices are unequal and differ by less than or equal to twice the maximum error. Thus, the largest approximation error for two Gittins indices can not result in a decision reversal.

[16] The realization of random variables between these simulations and the experiment is likely to be different. Online Appendix F reports alternative simulations based on the realization of reward rates and period lengths that subjects faced in the experiment. The predictions are largely robust to these differences.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 1. Predictions and results for each treatment

		Indefinite horizon		Finite horizon	
		2 Arms	3 Arms	2 Arms	3 Arms
(a) Predictions					
Experimentation	Overall	0.03	0.06	0.03	0.06
	Over time	0.05/0.04/0.02	0.10/0.08/0.04	0.05/0.02/—	0.10/0.04/—
Switching	Overall	0.11	0.13	0.13	0.14
	Over time	0.25/0.08/0.03	0.29/0.09/0.03	0.24/0.06/—	0.27/0.06/—
Best arm	Overall	0.82	0.72	0.80	0.69
	Over time	0.73/0.83/0.88	0.60/0.74/0.81	0.74/0.83/—	0.61/0.74/—
(b) Results					
Experimentation	Overall	0.18	0.25	0.19	0.30
	Over time	0.26/0.17/0.13	0.35/0.25/0.21	0.25/0.16/—	0.38/0.27/—
Switching	Overall	0.12	0.17	0.15	0.20
	Over time	0.20/0.10/0.06	0.31/0.13/0.07	0.20/0.12/—	0.29/0.14/—
Best arm	Overall	0.73	0.63	0.70	0.59
	Over time	0.64/0.74/0.79	0.50/0.65/0.73	0.64/0.74/—	0.50/0.64/—

Experimentation refers to non-myopic behavior when all arms have different expected immediate rewards. Switching refers to a different arm being chosen than in the previous round. Best Arm refers to the arm with the highest true underlying reward rate being chosen. The numbers given are the rate at which the event occurs. First-round decisions are excluded as they are uninformative. The three groupings for the "Over Time" rows refer to behavior in rounds 2–10, rounds 11–25, and after the 25th round.

We use these predictions to get a better understanding of optimal behavior in various multi-armed bandit problems. Our goal with these predictions is to guide our data analysis by creating a benchmark that can be compared to subject behavior rather than testing for comparative statics. We focus on three variables for these predictions: experimentation rate, switching rate, and best arm rate. Experimentation rate refers to how often a subject behaves non-myopically when all arms have different expected immediate rewards. Switching rate refers to how often a subject chooses an arm that is different from the arm that they previously chose. Best arm rate refers to how often a subject chooses the arm with the highest true underlying reward rate. We ignore first round decisions for each of these variables as first round behavior does not depend on the information generated from the bandit arms.

Table 1a) shows various predictions that are useful for our data analysis. This table shows that the experimentation rate is non-zero but still small. This is not unexpected as bandit arms often have quite different underlying reward rates. A related situation to experimentation is when subjects have to decide between arms that have the same expected reward rate but differing levels of information. Subjects, except for the last period of a finite horizon, are predicted to prefer the arm with less information if there are two arms with the same expected reward rate (Gittins et al., 2011). This arises as collecting information is valuable for the future. This is an important situation as it contains only exploration incentives and is thus a clean test of these incentives under optimal theory.

Table 1a) also predicts how behavior should change over time. It shows that subjects should experiment less over time, switch less over time, and be better at identifying the best arm over time. These predictions make sense as subjects have more information later in the period. Thus, over time, subjects have less incentive to experiment, should switch less due to more accurate beliefs, and should be better at identifying the best arm. In addition to these predictions, we also analyze switching based on

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

the last outcome (which is not shown in Table 1a). Subjects are predicted to never switch following a success, which is a property of these types of bandit problems (Robbins, 1952). The simulations also show that subjects should be less willing to switch, conditional on a failure, over time.

5. Results

In this section, we go over the results of the experiment. We first discuss the general results. We next discuss how behavior changes over the course of a bandit problem. We lastly discuss the implications of these results for subjects' strategies.

5.1. General results

Table 1b displays the general results of the experiment. This table suggests a few results. First, the experimentation and switching rates are greater than the optimal strategy predicts. Second, the best arm rate is lower than the optimal strategy predicts. Third, subjects appear to experiment less over time, switch less over time, and be better at identifying the best arm over time.

We now test the point predictions for the experiment. We test each point prediction in each treatment through a treatment-specific regression. Throughout the rest of this paper, regressions refer to random effects regressions with subject-level random effects and subject-level clustering (unless stated otherwise). First, we find significantly more experimentation than predicted at the 1% level in each treatment.[17] Second, we fail to reject the null hypothesis of the predicted amount of switching at the 10% level in each of our two-armed treatments. In each of the three-armed treatments, we find significant over-switching at the 5% level. Finally, we find that subjects choose the best arm less often than predicted at the 1% level in each treatment. These deviations from point predictions are important because they suggest that subjects do not use the optimal strategy. Result 1 summarizes these findings.

Result 1: *There is more experimentation than predicted by optimal theory in each treatment. There is more switching than predicted by optimal theory in the three-armed treatments. There is a lower best arm rate than predicted by optimal theory in each treatment.*

The over-experimentation in each treatment suggests that subjects may experiment differently than predicted by theory. In the predictions section, we mentioned how subjects are predicted to prefer arms that have less information when multiple arms have the same expected reward rates.[18] We analyze this prediction by focusing on the two-armed treatments but find similar behavior in the three-armed treatments. Subjects choose the arm with less information less than 20% of the time in the situation where it was not previously chosen. This provides further evidence that subjects experiment in a different way than predicted. We summarize this with Result 2.

Result 2: *Subjects do not always choose the arm with less information when there are two arms with the same expected reward rate.*

We now briefly discuss the comparative statics across treatments. When comparing the two-armed and three-armed indefinite horizons, we find more switching (p-value=0.046), more

[17] While decision noise would lead to over-experimentation, if subjects were otherwise playing optimally, this level of over-experimentation would require substantial noise. Simulations show subjects would need to tremble (i.e., choose randomly) 33%, 33%, 34%, and 42% of the time for the experimentation rates by treatment reported in Table 1b, respectively.

[18] One illustrative example is choosing between one arm that has one success and one failure and another with three successes and three failures. Both arms have an expected reward rate of one-half, but more information would be gained by pulling on the arm with fewer draws due to a larger shift in the estimate of the expected reward rate.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

experimentation (p-value=0.072), and a lower best arm rate (p-value=0.000) in the three-armed treatment. We find similar results for the finite horizon comparison but with some different significance levels (switching p-value=0.076, experimentation p-value=0.000, best arm p-value=0.000).[19] These results suggest that changing the number of arms leads to changes in our variables of interest that are consistent with the comparative statics from optimal theory. When comparing the two-armed indefinite and finite horizons, we find no significant differences (switching p-value=0.300, experimentation p-value=0.870, best arm p-value=0.164). When comparing the three-armed finite horizon to the three-armed indefinite horizon, we do find some significant differences (switching p-value=0.203, experimentation p-value=0.015, best arm p-value=0.008). However, we find no significant differences when comparing the two-armed (three-armed) finite horizon to the two-armed (three-armed) indefinite horizon once we control for the round. This suggests that the different round compositions between our finite and indefinite horizon treatments sometimes has an effect on aggregate results. It also suggests that we may see similar types of strategies across finite and indefinite horizons (controlling for the number of arms). The regressions in this paragraph are displayed in Online Appendix O.

5.2. Behavior over time

Table 1b suggests that subjects experiment less often, switch less often, and choose the best arm more often as time goes on. We test these suggestions through treatment regressions of the variable of interest on the round number. In each treatment, we find that subjects are significantly less likely to experiment in later rounds, less likely to switch in later rounds, and more likely to identify the best arm in later rounds. All of these results are significant at the 1% level. These results are important because they suggest that while subjects do not behave optimally, they make more targeted and better decisions over time. We summarize these results with Result 3.

Result 3: *The experimentation and switching rates decrease over time, while the best arm rate increases over time.*

We now look at switching rates, conditional on the last outcome, as we have predictions as to how these should evolve over time. Table 2 displays the switching rates over time following a failure and following a success. The table shows a few results. First, subjects sometimes switch following a success. However, switching rates are generally low following a success. Second, subjects generally switch at a higher rate following a failure than following a success. Additionally, subjects often stay on the arm they previously chose following a failure. Third, subjects tend to switch less over time (conditional on the last outcome). This suggests that subjects incorporate more than the previous outcome when making (switching) decisions. We summarize these observations with Result 4.

Result 4: *Subjects are more likely to switch following a failure than a success. Additionally, subjects are less likely to switch over time (conditional on the last outcome).*

One question that arises is whether subjects are switching more following a failure than a success because of the changes in incentives or because they overreact to the last outcome. It appears that subjects over-react to the last outcome. This is supported by treatment regressions of the decision to switch on an indicator variable for whether the last outcome was a failure and the number of successes and failures on each bandit arm. In these regressions, found in Online Appendix O, the indicator

[19] If we adjust for optimal predictions, we no longer find significant differences in switching and best-arm rates between two and three arms (all p-values>0.135). This suggests that subjects' underlying behavior may be similar in the presence of two and three arms.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 2. Switching based on last outcome

		Indefinite horizon		Finite horizon	
		Two-armed	Three-armed	Two-armed	Three-armed
Switch (Fail.):					
	Overall	0.21	0.31	0.27	0.37
	Over time	0.33/0.18/0.11	0.50/0.25/0.16	0.33/0.22/—–	0.47/0.30/—–
Switch (Succ.):					
	Overall	0.06	0.09	0.07	0.10
	Over time	0.10/0.05/0.03	0.18/0.07/0.04	0.10/0.05/—–	0.16/0.06/—–

"Switch (Fail.)" refers to the switching rate following a failure. "Switch (Succ.)" refers to the switching rate following a success. The three groupings for the "Over time" rows refer to behavior in rounds 2–10, rounds 11–25, and after the 25th round.

variable for the last outcome being a failure is positive and significant at the 5% level. Thus, subjects' response to the previous outcome is driven by more than the change in incentives. We summarize this result with Result 5.

Result 5: *Subjects are more likely to switch following a failure than a success after conditioning for the failures and successes on each arm.*

5.3. Summarizing experimental results

The results from the earlier subsections have a few implications for estimated strategies. First, subjects do not appear to use the optimal strategy. This is shown by various results: (i) subjects finding the ex ante best arm less often than predicted, (ii) subjects experimenting in a different way than predicted, and (iii) subjects treating failures differently than successes after controlling for the number of successes and failures on each arm. Second, while subjects do not appear to use the optimal strategy, they do appear to be making better decisions over time. The dynamics of subject behavior within a bandit problem are consistent with incentives and exhibit a similar pattern of behavior to what optimal theory suggests. Third, subjects appear to focus on more than just the last outcome when making decisions. This suggests that while the last outcome may be important, subjects avoid strategies that make decisions solely based on the last outcome.

6. Strategies

This section investigates the possible strategies that subjects use in multi-armed bandit problems. The first two subsections introduce and estimate commonly suggested deterministic and probabilistic strategies.[20] While some of these strategies do not capture the aggregate behavioral trends identified in the previous section, it is possible that some subjects use them. In the third subsection, we build on some of the commonly suggested strategies using the behavioral trends from the experiment and compare these new strategies to our previously estimated strategies. In the fourth subsection, we run some robustness checks on our results.

Throughout this section, for succinctness, we only briefly describe the previously suggested strategies that we estimate. We provide more details on each of these strategies in Online Appendix H.

[20]The term "deterministic" refers to strategies that typically predict a unique arm and the term "probabilistic" refers to strategies that typically predict a non-zero probability of choosing any individual arm.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 3. List of commonly suggested deterministic strategies

Name	Description	Versions
Optimal	Choose optimal action	
Myopic	Choose arm with highest immediate expected reward rate	• Correct • Empirical
Hot-hand-N	Switch if arm pulled at least *N* consecutive times and resulted in *N* consecutive losses	• *N*:1-5
Never Switch	Stay on previous arm	• Copy Last • Copy First
Exponential smoothing	Choose arm with highest E-S index $(\gamma I_{success} * 0.50 + (1-\gamma)ES_i(t-1))$	
Last-N	Choose arm with highest expected reward rate (based on last N draws in each arm)	• Bayesian (*N*:1-5) • Empirical (*N*:1-5)
Simple	Choose arm with largest simple index $(\omega^i_{t-1} - l^i_{t-1} * D)$	

6.1. Deterministic strategies

We consider seven types of commonly suggested deterministic strategies.[21] These strategies are displayed in Table 3. The first two strategies are the optimal and myopic strategies. The optimal strategy predicts that a subject always chooses the optimal action as described in subsections 2.2 and 2.3. The myopic strategy predicts that a subject always chooses the arm with the highest immediate expected payoff. We analyze two different versions of the myopic strategy. The first version is Myopic-Correct, where subjects' decisions are based on the correct calculation of the immediate expected reward. The second version is Myopic-Empirical, where subjects' beliefs of an arm's expected immediate reward are equivalent to that arm's average payout.[22] We add this version as subjects may use empirical reward rates to avoid the cognitive costs of updating.

The next two strategies are the hot-hand-N and never-switch strategies. The hot-hand-N strategy is where a subject switches arms once an arm has been pulled at least N consecutive times and has resulted in N consecutive failures. In the case that a subject switches, they switch randomly to one of the other arms. We estimate the hot-hand-1 through hot-hand-5 strategies.[23] The fourth strategy is the never-switch strategy, where subjects are predicted to not switch. We estimate a copy-last version (where subjects choose the same arm as they selected in the last round) and a copy-first version (where subjects choose the arm that they selected in the first round). Although these strategies result in the same predictions in theory, they result in different predictions when subjects tremble.

The next two strategies are the exponential-smoothing and last-N strategies. In the exponential-smoothing strategy, subjects choose the arm with the largest exponential-smoothing index. The initial index ($ES_i(1)$) for each arm is 0.25, which is the ex ante expected immediate payoff of each arm. In each round t, where a subject previously samples from arm i, the index $ES_i(t) = \gamma I_{success} * 0.50 + (1-\gamma)ES_i(t-1)$, where $0 \leq \gamma \leq 1$ (non-chosen arms have unchanged indices) and 0.50 is the payoff

[21] The optimal, myopic, hot-hand, exponential-smoothing, last-N, and simple strategies were estimated in Gans et al. (2007). We add empirical versions of the Myopic and last-N strategies. Copy-Last is equivalent to $HH-\infty$, which fits well with Gans et al. (2007). We add Copy-First as a possible alternative.

[22] The initial expected immediate reward is indeterminate before a lever is pulled (because beliefs should be based on empirical results). We treat this value as a free parameter. We estimate this strategy for various levels of this free parameter ({0, ..., .995, 1}).

[23] We do not go higher than hot-hand-5 because these strategies become similar to our never-switch strategies. It is rare that a subject is on an arm for six or more consecutive failures.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

in the case of a success.[24] The sixth strategy is the Last-N strategy, which has two versions. The first version is Last-N-Empirical, where subjects' beliefs of an arm's expected immediate reward are that arm's average payout from the last N draws. The second version is Last-N-Bayesian, where subjects Bayesian update each arm's expected reward using that arm's last N draws and the initial prior. For each version, we estimate five Last-N strategies ($N \in \{1, 2, 3, 4, 5\}$).

The last strategy is the simple strategy. The simple strategy is an index strategy where the index value of each arm is given by $\omega^i_{t-1} - l^i_{t-1} * D$, where i denotes the arm, ω^i_{t-1} denotes the number of previous successes on arm i, l^i_{t-1} denotes the number of previous losses on arm i, and D denotes the weight on losses.[25]

We estimate the best-fitting deterministic strategies in the following way. For each subject, we obtain the log-likelihood for each strategy.[26] In the two-arm treatments, we model the log-likelihood for these strategies as

$$LL = log\left(\prod_P \prod_R \left((1-\beta) + \frac{\beta}{2}\right)^{I_{CU}} \left(\frac{\beta}{2}\right)^{I_{IU}} \left(\frac{1}{2}\right)^{I_{Tie}}\right),$$

where $\prod_P$ multiplies over each period and $\prod_R$ multiplies over each round. The indicator variable I_{CU} refers to a unique correct prediction, the indicator variable I_{IU} refers to a unique incorrect prediction, and the indicator variable I_{Tie} refers to a prediction that could justify drawing from either box.[27] The parameter β can be interpreted as a tremble where a subject loses concentration and behaves uniformly random.[28] We use BIC to compare different strategies as it penalizes models for the number of free parameters.[29]

In the three-arm treatments, the log-likelihood for these strategies is modeled as

$$LL = log\left(\prod_P \prod_R \left((1-\beta) + \frac{\beta}{3}\right)^{I_{CU}} \left(\frac{(1-\beta)}{2} + \frac{\beta}{3}\right)^{I_{C2}} \left(\frac{(\beta)}{3}\right)^{I_I} \left(\frac{1}{3}\right)^{I_{Tie}}\right).$$

This is analogous to how the log-likelihood is modeled in the two-arm treatments. Once again, the indicator variable I_{CU} refers to a unique correct prediction. The indicator variable I_{C2} refers to a correct prediction when the strategy predicted either of two arms. The indicator variable I_I refers to the strategy not predicting the current choice. The indicator variable I_{Tie} refers to a prediction that could justify drawing from any box.

The results of the deterministic strategy estimation are presented in Table 4. This table displays the number of subjects that each deterministic strategy fits best (among only deterministic strategies). We uncover a few insights from this comparison. We find that most subjects can be best classified (among deterministic strategies) as using a never-switch strategy or a hot-hand strategy. These strategies

[24]Due to identifiability issues, we estimate many versions of this strategy ($\gamma \in \{0, 0.005.0.010, 0.015, \ldots, 0.990, 0.995, 1\}$). When estimating index strategies, identifiability issues arise as a small change in a parameter will often have no effect on the ranking of indices. We use a grid search method to address this (this is the same approach as used in Gans et al. (2007)).

[25]Due to identifiability issues, we estimate $D \in \{0, 0.005, 0.010, 0.015, \ldots, 2.995, 3\}$. A simple strategy with $D > 3$ does not outperform the strategies within this range for any subject.

[26]We prefer this approach over a finite mixture model as it allows us to estimate each strategy at the individual level. Furthermore, a finite mixture model is not computationally tractable in this situation due to the number and complexity of the strategies that we consider.

[27]When a deterministic strategy predicts multiple arms, we assume subjects choose one of these arms randomly. For example, multiple arms can tie for the highest expected reward rate in the myopic strategy (e.g., each arm has one success and zero failures).

[28]We utilize this type of tremble because it easily extends to bandit problems with more arms.

[29]We also use AIC, which leads to similar results and can be found in Online Appendix J. BIC places a larger penalty on free parameters than AIC, given the number of observations we have. We prefer BIC over AIC for this reason. We provide more details on BIC and AIC in Online Appendix N.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 4. Comparison of deterministic strategies

	N-S	HH	Myopic	Last-N	Simple	Optimal	ES
Two indefinite	27	9	5	6	6	1	0
Two finite	23	13	9	7	1	0	0
Three indefinite	17.5	12.5	15	4	4	0	1
Three finite	17	16	5	13	2	1	0

Displays the number of subjects each deterministic strategy fits best (among only deterministic strategies). If multiple strategies fit a subject equally, each strategy is given equal weight.

Table 5. List of commonly suggested probabilistic strategies

Name	Description	Versions
Epsilon-greedy	Behave myopically (randomly) with prob. $1-\omega_0 e^{\omega_1(t-1)}$ ($\omega_0 e^{\omega_1(t-1)}$)	• Correct • Empirical
Epsilon-first	Behave myopically (randomly) with probability $1-\epsilon_1$ (ϵ_1) in first N rounds; behave myopically (randomly) with probability $1-\epsilon_2$ (ϵ_2) otherwise	• Correct (*N*:2-10) • Empirical (*N*:2-10)
Win-stay lose-shift	Switch with probability $\gamma_{success}$ following a success and $\gamma_{failure}$ following a failure	
Randomization	Each arm has equal probability of being chosen	
Thompson sampling	Draw probability of arm success from posterior distribution and choose arm with highest draw	
Reinforcement learning	Each arm's probability of being chosen is a function of propensity weights	• Cumulative • Discounted
Myopic QR	Each arm's probability of being chosen is based on a logit form of expected rewards	• Correct • Empirical

suggest that the last arm chosen has an important role in subjects' decision making.[30] Additionally, strategies that compare immediate expected rewards (myopic and last-N) fit more subjects than the optimal strategy.

6.2. Probabilistic strategies

We now consider seven different commonly suggested probabilistic strategies.[31] These strategies are displayed in Table 5. The first two strategies are the epsilon-greedy and epsilon-first strategies. In the epsilon-greedy strategy, subjects behave myopically with probability $1-\epsilon$ and behave randomly with probability ϵ. We allow for ϵ to decrease over time as we set $\epsilon = \omega_0 e^{\omega_1(t-1)}$, where t is the current round number. In the epsilon-first strategy, subjects in the first N rounds behave myopically with probability $1-\epsilon_1$ and behave randomly with probability ϵ_1. After the first N rounds, subjects

[30] The vast majority of subjects identified as never-switch are best fit by the copy-last version.

[31] Although their original references are unclear, the epsilon-greedy and epsilon-first strategies are often suggested (Slivkins, 2021). Probabilistic win-stay lose-shift is explored in Hu et al. (2013). Randomization is a special case of this strategy. Thompson Sampling was originally suggested in early work on the multi-armed bandit problem (Thompson, 1933, 1935). Reinforcement learning is a common strategy in economics (based on Roth & Erev (1995)) and psychology (based on Rescorla & Wagner (1972)). Myopic QR is a popular way of generally behaving myopically (through the softmax function) in computer science (Sutton & Barto, 1998). We also estimate the EXP3 strategy, which is suggested in Auer et al. (2002). We leave details on this strategy to Online Appendix H as it does not fit any subjects best and is complex.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

behave myopically with probability $1-\epsilon_2$ and behave randomly with probability ϵ_2. We estimate nine epsilon-first strategies ($N \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$).

The next two strategies are the probabilistic win-stay lose-shift and randomization strategies. In the probabilistic win-stay lose-shift strategy, a subject switches arms with probability $\gamma_{success}$ following a success and with probability $\gamma_{failure}$ following a failure. In the case that a subject switches, they switch randomly to one of the other arms. In the randomization strategy, each subject plays each action with equal probability.

The next strategy is the Thompson Sampling strategy. In the Thompson Sampling strategy, subjects maintain a current posterior of beliefs over the possible success probabilities for each arm, draw a probability of success from each posterior, and then choose the arm with the highest drawn probability of success.

The last two strategies are reinforcement learning and the Myopic QR strategy. We estimate two versions of reinforcement learning. In the first version (Roth & Erev, 1995), we model reinforcement learning as subjects having an initial propensity for each arm and subjects increasing the propensity for a given arm by the reward it obtains when chosen.[32] The probability that each arm is chosen is obtained by dividing each arm's propensity over the sum of all the arms' propensities. In the second version (Rescorla & Wagner, 1972), subjects' propensity is similar to the exponential-smoothing index except that the initial index/propensity is now a free parameter. The probability that an arm is chosen is a logit function of the propensities. In the Myopic QR strategy, subjects choose each action based on a logit function (similar to Quantal Response Equilibrium) that incorporates each arm's expected immediate reward. We estimate one version of this strategy using the arms' correct expected immediate reward and another using the empirical average as a proxy for the expected immediate reward.

Our estimation for probabilistic strategies is complicated by the differing nature of these strategies. One simple way of describing the way we estimate these strategies is that we add $log(p)$ to the log-likelihood after every choice, where p is the probability that the specific strategy places on that action being chosen. Thus, we model the log-likelihood for these strategies as

$$LL = log\left(\prod_{P}\prod_{R}(p_1)^{I_{Arm1}}(p_2)^{I_{Arm2}}(p_3)^{I_{Arm3}}\right),$$

where (p_i) denotes the probability that the given strategy assigns to choosing arm i (given the bandit history). Online Appendix H displays more information about how the log-likelihoods for these strategies are developed. It is not necessary to add a tremble β for these estimations as these strategies are nondeterministic.[33]

The results of the probabilistic strategy estimation are presented in Table 6, which considers all previously suggested strategies (i.e., deterministic strategies are included).[34] An overwhelming number of subjects are best fit by either the probabilistic win-stay lose-shift strategy or reinforcement learning. Probabilistic win-stay lose-shift best fits the most subjects, but reinforcement learning is a close second. The prevalence of the probabilistic win-stay lose-shift strategy further suggests that the last arm plays an important role in decision making but also suggests that subjects are sensitive to the last outcome. This table also shows that very few subjects are best fit by deterministic strategies. In each treatment, fewer than seven subjects are best fit by a deterministic strategy.

[32]The initial propensity is a free parameter.

[33]We choose not to add a tremble β to the probabilistic strategies because it can reduce the identifiability of the other estimated parameters for these probabilistic strategies (additionally, some strategies also have a built-in tremble). This leads to a slight difference in how probabilistic and deterministic strategies are estimated. One way of thinking about the estimations from the previous subsection is to think of them as partially deterministic strategies. This difference in estimation approaches is unimportant as our 'probabilistic' strategies fit much better than our 'deterministic' strategies.

[34]Online Appendix B shows that behavior appears to stabilize in the last 20 periods. Similar results are found when using the last 20 periods instead of the entire dataset.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 6. Comparison of previously suggested strategies

	WSLS	RL	M. QR	N-S	HH	TS	Last-N	ϵ-G
Two indefinite	29	15	6	3	0	0	1	0
Two finite	27	15	5	2	4	0	0	0
Three indefinite	24	21	4	1	1	1	1	1
Three finite	29	15	4	2	2	2	0	0

Displays the number of subjects each strategy fits best in each treatment. If multiple strategies fit a subject equally, each strategy is given equal weight toward fitting a subject best.

It is interesting that the probabilistic win-stay lose-shift strategy best fits a majority of subjects in Table 6. The previous section suggested that subjects deviate from this strategy by switching less over time (conditional on the last outcome). In Online Appendix A, we show that even subjects who are classified as using probabilistic win-stay lose-shift in Table 6 switch less over time (conditional on the last outcome). Thus, there may be strategies that better capture subject behavior. In the next subsection, we use both the data moments from the experiment and the best-fitting strategies from this subsection to uncover better-fitting strategies.

6.3. Additional strategies

We design three new strategies in order to try to better uncover subject behavior. These three new strategies are decreasing win-stay lose-shift, biased reinforcement learning, and biased myopic. Each of these new strategies is an adaptation of one of the three previously best fitting strategies.

The decreasing win-stay lose-shift strategy adapts the probabilistic win-stay lose-shift strategy by allowing for subjects to be less willing to switch, conditional on the last outcome, over time. We estimate this strategy because Result 4 shows that subjects appear to switch less (conditional on the last outcome) over time. In this strategy, the probability that an individual switches following a failure is given by $\omega_{0F}e^{\omega_{1F}(t-1)}$ and the probability that an individual switches following a success is given by $\omega_{0S}e^{\omega_{1S}(t-1)}$. The variable t is the round number. We restrict $0 \leq \omega_{0F} \leq 1, 0 \leq \omega_{0S} \leq 1, \omega_{1F} \leq 0$, and $\omega_{1S} \leq 0$.

The biased reinforcement learning strategy adapts the Rescorla & Wagner (1972) reinforcement model by penalizing arms that require switching. In this new model, an individual's propensity updates, when previously chosen, by the following equation:

$$V_t = (1-\alpha)V_{t-1} + \alpha R_{t-1},$$

where R_{t-1} is the reward obtained and $0 \leq \alpha \leq 1$. The probability that an individual stays on arm i is then given by

$$\frac{e^{\lambda V_{t,i}}}{e^{\lambda V_{t,i}} + \sum_{j \neq i} e^{\lambda (c+V_{t,j})}},$$

where $c \leq 0$ and $\lambda \geq 0$.[35] This strategy is essentially reinforcement learning but is biased toward the previously chosen arm.

We impose this bias into reinforcement learning (and myopic QR) for two main reasons.[36] First, the previously estimated strategies that tend to fit best either (i) treat previously selected and unselected arms differently (probabilistic win-stay lose-shift) or (ii) suggest that subjects are influenced by all bandit outcomes (reinforcement learning and myopic QR). We may be able to build a better-fitting strategy by incorporating these two elements. Second, subjects treat previously chosen arms

[35] Given the logit function, this is equivalent to a function where the previously chosen arm gets its propensity adjusted upwards by $-c \geq 0$.

[36] Online Appendix L shows that our biased strategies can capture the win-stay lose-shift violations.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 7. Comparison of previously suggested and new strategies

	B. RL	B. M	RL	DWSLS	M. QR	WSLS	HH	N-S	TS	ϵ-G
Two indefinite	17	23	8	3	2	1	0	0	0	0
Two finite	26	17	2	2	2	3	1	0	0	0
Three indefinite	19	22	6	3	1	0	1	1	0	1
Three finite	29	15	4	2	1	1	0	1	1	0

Displays the number of subjects each strategy fits best in each treatment after incorporating the new strategies. If multiple strategies fit a subject equally, each strategy is given equal weight toward fitting a subject best.

differently from previously unselected arms. Subjects are more likely to choose an arm, conditional on the number of failures and successes on each arm, if they previously chose it.[37] This is true even when the last outcome was a failure.

The biased myopic strategy adapts the myopic QR strategy by penalizing arms that have not been previously chosen. Under this strategy, the probability that an individual stays with the current arm i is equal to

$$\frac{e^{\lambda(I_{success}*\alpha+E[reward_i])}}{e^{\lambda(I_{success}*\alpha+E[reward_i])}+\sum_{j\neq i}e^{\lambda(c+E[reward_j])}},$$

where $I_{success}$ is an indicator variable denoting a previous success, α is an upward adjustment for a previous success, and c is a cost for switching. We restrict $\alpha \geq 0$ and $c \leq 0$. The idea behind this strategy is that subjects respond to the expected reward of each arm but require a premium to switch to another arm.[38] We impose a bias toward the previously selected arm (through c) into myopic QR for a similar reason as reinforcement learning. The presence of α allows subjects' bias toward the previous arm to depend on the previous outcome (biased reinforcement learning has something similar through δ).

Table 7 displays the best-fitting strategies among the previously estimated strategies and the three new strategies for each treatment. There are a few takeaways from this table. First, a slight majority of subjects can be classified as using some sort of reinforcement learning. Second, an overwhelming majority of subjects can be classified as being biased toward their previous choice. Third, probabilistic win-stay lose-shift does not do very well once we have incorporated biases into reinforcement learning and myopic QR.

6.4. Robustness checks

In the previous subsection, we found that our biased strategies best fit the most subjects compared to previously suggested bandit strategies. However, it is possible that commonly suggested strategies from outside of the bandit literature may do well once adapted to our environment.[39] Additionally, it is possible that our estimation would result in different conclusions if conducted on a dataset that did not inform our strategy design. In this subsection, we briefly discuss these two robustness checks.

We first structurally estimate two successful strategies from outside of the bandit literature to test whether they can better fit subjects than our biased strategies. The first strategy is a modified I-SAW

[37] This is supported by regressions, for each arm and treatment, of an indicator variable for whether the subject chose arm i on the number of failures and successes on each arm and an indicator variable for whether the subject last chose arm i. These regressions can be found in Online Appendix O.

[38] We estimate both Correct and Empirical versions of this strategy.

[39] We thank an anonymous referee for this suggestion and for a suggestion that a selection criterion that does not rely on log-likelihood may result in different strategies performing better. We show that our estimation is robust to different selection criteria in Online Appendix J.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Table 8. Comparison of strategies in Robustness treatment

	B. RL	B. M	RL	M. QR	WSLS	N-S	TS	LastN	EWA	I-SAW
Original set	26	19	5	3	2	2	1	1	—	—
EWA & I-SAW	25	18	5	3	2	2	1	1	2	0

Displays the number of subjects each strategy fits best in the Robustness treatment. If multiple strategies fit a subject equally, each strategy is given equal weight towards fitting a subject best. "Original set" refers to estimating all of the strategies except for EWA and I-SAW. "EWA & I-SAW" refers to estimating every strategy that was previously estimated in the paper.

model that is based on Nevo & Erev (2012). As mentioned in the Introduction, Nevo & Erev (2012) introduce I-SAW, which is a type of reinforcement learning model that predicts behavior well in decisions from experience. The original I-SAW model is not directly applicable to our environment, and we thus have to modify it. The second strategy is the Experience-Weighted Attraction (EWA) model.[40] We estimate two versions of this model: the original version (Camerer & Ho, 1999) and the self-tuning version (Ho et al., 2007). Both of these versions are reinforcement learning models and we adapted both to our environment. Online Appendix H displays the details of these strategies and how they were adapted. Table 17 in Online Appendix I displays the results of the estimation once we add these strategies. In our experiment, we find that only one subject is best fit by the modified I-SAW strategy, and only six subjects are best fit by a modified EWA strategy. There are still 164 subjects (76.3%) who are best fit by one of our biased strategies. This estimation shows that our biased strategies are able to capture subject behavior in a way that the other strategies cannot.

We lastly consider the possibility that our biased strategies only performed well in our estimations because we estimated them on the same subjects that inspired them. It is possible that we designed strategies that fit well for this specific set of subjects and that these strategies may not perform well under a different set of subjects. This concern is mitigated by our relatively large sample of subjects and by the biased strategies fitting well across four different environments. However, in July 2024, we conducted a new treatment to test this possibility. This "Robustness" treatment is similar to our three-armed finite horizon bandit treatment, but now a bandit problem lasts 18 rounds (instead of 25 rounds). We additionally had subjects face 40 bandit problems to keep the overall number of decisions similar to our previous experiment. We chose a shorter bandit problem as we additionally wanted to test whether strategies with inertia would still perform well in a shorter bandit problem. Details on the Robustness treatment can be found in Online Appendix M.

Table 8 displays the number of subjects best fit by each strategy in the Robustness treatment. We observe very similar results to our original experiment. In the Robustness treatment, 72.9% of subjects are best fit by a biased strategy. This is very similar to the 76.3% of subjects that are best fit by a biased strategy in the original data. Additionally, we once again find a slight majority of subjects best fit by a reinforcement learning model. The results of this estimation are qualitatively consistent with out-of-sample exercises in Table 25 and 26 in Online Appendix M. Table 25 shows that behavioral predictions based on the best-fitting strategies in the original experiment predict aggregate subject behavior very well in the Robustness treatment. Table 26 reports the best-fitting strategies in the Robustness treatment when the strategy parameters are determined from the original experiment. Our biased strategies in this exercise capture a majority of subjects. Overall, our results suggest that our biased strategies capture behavior well across many different environments and many different subjects.

[40]EWA is used as a base for a model that fits well (Ferecatu & De Bruyn, 2022) for a bandit based on a normal distribution. The full model in Ferecatu & De Bruyn (2022) is not suitable for our specific estimation approach due to its large number of free parameters.

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

7. Discussion

In this paper, we conduct an exploratory analysis on behavior in the classic multi-armed bandit problem. We estimate fourteen different strategies from the multi-armed bandit literature and find that most subjects are best fit by either a probabilistic win-stay lose-shift strategy or reinforcement learning. However, we show that it is unlikely that subjects are using these strategies. We develop three new strategies and show that most subjects are best fit by either a biased reinforcement learning strategy or a biased myopic strategy. In this discussion section, we discuss the implications of these results.

Our results have various important takeaways. The first takeaway is that most subjects appear to judge bandit arms through a reinforcement learning process. We find that a majority of subjects in our initial experiment (112/215) and in our Robustness treatment (32/59) are best fit by a strategy that incorporates reinforcement learning. This is impressive given that the large number of strategies that were estimated and that BIC places a larger penalty on strategies with more parameters. The success of reinforcement-like strategies suggests that most subjects do play relatively successful arms more often but do not explicitly evaluate these arms based on a correct Bayesian updating process.

The second takeaway is that subjects tend to place a premium on the last chosen arm, given how they evaluate arms. This is shown by the overwhelming majority of subjects in both our initial experiment (164/215) and Robustness treatment (43/59) that are best fit by a biased strategy. Biased reinforcement learners evaluate arms through a reinforcement learning process, but are more likely to choose their previous arm than their propensities predict. Similarly, biased myopic subjects evaluate arms through a myopic process but are more likely to choose their last chosen arm than the expected reward rate of each arm predicts. The success of these strategies suggest that subjects are more likely to play their previous action than their underlying evaluation process suggests.

The third takeaway is that subjects experiment differently than theory predicts. This is suggested by the relatively poor fit of the optimal strategy. Our strategy estimation instead suggests that experimentation occurs through randomness and subjects staying on myopically suboptimal arms. Randomness is consistent with the fit of the biased strategies, which allow for intelligent noise through logit functions. This intelligent noise is more sophisticated than an arbitrary "tremble," as it is proportional to each arm, meaning that randomness is more likely when arms are evaluated as being more similar. We thus expect subjects to experiment more often when bandit arms have returned rewards at similar rates as subjects appear to evaluate arms using a reward-based process that is correlated with expected immediate reward. Subjects experimenting through staying on myopically suboptimal arms follows from our best-fitting strategies. Simulations on subjects' implied strategies show that shutting off or turning down either of these channels in isolation reduces the strategy's average payoffs. These channels that can induce experimentation prove to be beneficial, given how subjects otherwise make decisions.

The fourth takeaway is that estimations should focus on strategies that both exhibit inertia and that have probabilistic choice based on perceived similarity in the bandit arms. In terms of the previously suggested strategies that we estimate, most of them lacked one of these two considerations. For example, while the probabilistic win-stay lose-shift strategy initially performed well, it suffered from only focusing on the previous outcome. Subjects respond to more than just the previous outcome and appear to condition their behavior based on reward-based evaluations of the arms. Additionally, strategies that focus solely on the observed bandit outcomes suffer from not incorporating subjects' tendency to treat arms differently based on whether they were last chosen. Only our two biased strategies (out of the strategies that were estimated) directly modeled both of these features. It thus seems fruitful to focus on strategies that model both inertia and probabilistic choice based on perceived similarity in the bandit arms.

There are many paths for future research. Future research could try to uncover behavioral factors that influence experimentation in multi-armed bandit problems. It would be interesting to uncover why subjects treat arms differently based on whether they were last chosen. Future research could also

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

analyze behavior in different types of multi-armed bandit environments. There are many extensions of the classic multi-armed bandit problem that would be interesting to analyze, such as restless bandits and multi-armed bandits that include safe options. Lastly, future research could use other methods to try to uncover subject behavior. We focus on estimating strategies, but other papers could try to elicit strategies more directly by allowing subjects to build their own strategies.

Acknowledgments. We are grateful for helpful comments by Jason Aimone, Tim Cason, Braxton Gately, Nicolas Klein, Yaroslav Rosokha, Zach Ward, and conference and seminar audiences at Baylor University, Purdue University, the University of Toronto, the CREED/TI workshop on Experimentation: Learning and Information Design, the SEA Meetings in Houston, and the ESA Meetings in Tucson and Bologna. We also thank the editor, an anonymous coeditor, and two anonymous referees. The replication material for the study is available at https://doi.org/10.17605/OSF.IO/4PW7M.

References

Addicott, M. A., Pearson, J. M., Wilson, J., Platt, M. L., & McClernon, F. J. (2013). Smoking and the bandit: A preliminary study of smoker and nonsmoker differences in exploratory behavior measured with a multi-armed bandit task. *Experimental and Clinical Psychopharmacology*, *21*(1), 66–73.

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

Anderson, C. (2012). Ambiguity aversion in mutli-armed bandit problems. *Theory and Decision*, *72*, 15–33. https://doi.org/10.1007/s11238-011-9259-2

Arechar, A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, *21*, 99–131. https://doi.org/10.1007/s10683-017-9527-2

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*(1), 48–77. https://doi.org/10.1137/S0097539701398375

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory*, *10*, 55–77. https://doi.org/10.1007/s001990050146

Banovetz, J., & Oprea, R. (2023). Complexity and procedural choice. *American Economic Journal: Microeconomics*, *15*(2), 384–413.

Camerer, C., & Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Chen, D., Schonger, M., & Wickens, C. (2016). oTree: An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*(2214–6350), 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Dal Bó, P., & Fréchette, G. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, *101*, 411–429. https://doi.org/10.1257/aer.101.1.411

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879. https://doi.org/10.1038/nature04766

Deck, C., & Kimbrough, E. (2017). Experimenting with contests for experimentation. *Southern Economic Journal*, *84*(2), 391–406. https://doi.org/10.1002/soej.12185

Duffy, J., & Feltovich, N. (1999). Does observation of others affect learning in strategic environments? An experimental study. *International Journal of Game Theory*, *28*, 131–152. https://doi.org/10.1007/s001820050102

Erev, I., & Haruvy, E. (2016). Learning and the economics of small decisions. In Kagel, J. H. & Roth, A. E., (Eds.), *The handbook of experimental economics*. Princeton University Press, *2*, 638–716.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, *88*(4), 848–881.

Feltovich, N. (2000). Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. *Econometrica*, *63*(3), 605–641. https://doi.org/10.1111/1468-0262.00125

Ferecatu, A., & De Bruyn, A. (2022). Understanding managers' trade-offs between exploration and exploitation. *Marketing Science*, *41*(1), 139–165.

Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, *102*(2), 720–749. https://doi.org/10.1257/aer.102.2.720

Gans, N., Knox, G., & Croson, R. (2007). Simple models of discrete choice and their performance in bandit experiments. *Manufacturing & Service Operations Management*, *9*(4), 383–408.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277–286.

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, *7*(3), 391–415.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 148–164. https://doi.org/10.1111/j.2517-6161.1979.tb01068.x

Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Gittins, J. C., Glazebrook, K., & Weber, R. (2011). *Multi-armed bandit allocation indices* (2nd ed.). John Wiley & Sons.
Gittins, J. C., & Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J. (Ed.), *Progress in statistics* (pp. 241–266). North-Holland Publishing Company.
Ho, T.-H., Camerer, C., & Chong, J.-K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, *133*(1), 177–198.
Hoelzemann, J., & Klein, N. (2021). Bandits in the lab. *Quantitative Economics*, *12*(3), 1021–1051. https://doi.org/10.3982/QE1389
Horowitz, A. (1973). *Experimental study of the two-armed bandit problem*. Unpublished doctoral dissertation, University of North Carolina.
Hotaling, J. M., Navarro, D. J., & Newell, B. R. (2021). Skilled bandits: Learning to choose in a reactive world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(6), 879–905.
Hu, Y., Kayaba, Y. & Shum, M. (2013). Nonparametric learning rules from bandit experiments: The eyes have it! *Games and Economic Behavior*, *81*, 215–231. https://doi.org/10.1016/j.geb.2013.05.003
Hudja, S. (2019). *Voting for experimentation: A continuous time analysis*. *Unpublished Manuscript*. https://dx.doi.org/10.2139/ssrn.3473426
Hudja, S. (2021). Is experimentation invariant to group size? A laboratory analysis of innovation contests. *Journal of Behavioral and Experimental Economics*, *91*, 10166. https://doi.org/10.1016/j.socec.2020.101660
Hudja, S., & Woods, D. (2024). Exploration versus exploitation: A laboratory test of the single-agent exponential bandit model. *Economic Inquiry*, 62(1), 267–286. https://doi.org/10.1111/ecin.13164
Kip Viscusi, W., & DeAngelis, S. (2018). Decision irrationalities involving deadly risks. *Journal of Risk and Uncertainty*, *57*(3), 225–252.
Meyer, R., & Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, *41*(5), 817–834. https://doi.org/10.1287/mnsc.41.5.817
Navarro, D. J., Tran, P., & Baz, N. (2018). Aversion to option loss in a restless bandit task. *Computational Brain & Behavior*, *1*, 151–164. https://doi.org/10.1007/s42113-018-0010-8
Nevo, I., & Erev, I. (2012). On surprise, change, and the effect of recent outcomes. *Frontiers in Psychology*, *3*, 1–9. https://doi.org/10.3389/fpsyg.2012.00024
Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.
Robbins, H. (1952). Some aspects of the sequential design of experiments. *American Mathematical Society*, *58*(5), 527–535. https://doi.org/10.1090/s0002-9904-1952-09620-8
Romero, J., & Rosokha, Y. (2018). Constructing strategies in the indefinitely repeated prisoner's dilemma game. *European Economic Review*, *104*, 185–219. https://doi.org/10.1016/j.euroecorev.2018.02.008
Romero, J., & Rosokha, Y. (2023). Mixed strategies in the indefinitely repeated prisoner's dilemma. *Econometrica*, *91*(6), 2295–2331. https://doi.org/10.3982/ECTA17482
Rosokha, Y., & Younge, K. (2020). Motivating innovation: The effect of loss aversion on the willingness to persist. *Review of Economics and Statistics*, *102*(3), 569–582. https://doi.org/10.1162/rest_a_00846
Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*(1), 164–212. https://doi.org/10.1016/S0899-8256(05)80020-X
Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, 101261. https://doi.org/10.1016/j.cogpsych.2019.101261
Slivkins, A. (2021). *Introduction to multi-armed bandits*. Unpublished Manuscript.
Snowberg, E., & Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, *111*(2), 687–719. https://doi.org/10.1257/aer.20181065
Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367.
Sutton, R., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.
Thompson, W. (1933). On the likelihood that one unknown probability exceed another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294. https://doi.org/10.2307/2332286
Thompson, W. (1935). On the theory of apportionment. *American Journal of Mathematics*, *57*(2), 450–456. https://doi.org/10.2307/2371219
Toyokowa, W., Kim, H.-R., & Kameda, T. (2014). Human collective intelligence under dual exploration-exploitation dilemmas. *PLOS One*, *9*(4), e95789.
von Helversen, B., Mata, R., Samanez-Larkin, G. R., & Wilke, A. (2018). Foraging, exploration, or search? on the (lack of) convergent validity between three behavioral paradigms. *Evolutionary Behavioral Sciences*, *12*(3), 152–162.
Wang, Y. (1997). Error bounds for calculation of the gittins indices. *Australian Journal of Statistics*, *39*(2), 225–233. https://doi.org/10.1111/j.1467-842X.1997.tb00538.x
Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. bioRxiv. https://doi.org/10.1101/106286

https://doi.org/10.1017/eec.2025.10027 Published online by Cambridge University Press

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behavior*, *2*, 915–924. https://doi.org/10.1038/s41562-018-0467-4

Yi, S., Steyvers, M., & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, *2*(2), 81–101.

Cite this article: Hudja, S., & Woods, D. (2025). Strategies in the multi-armed bandit. *Experimental Economics*, 1–23. https://doi.org/10.1017/eec.2025.10027