

Marketing Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Understanding Managers' Trade-Offs Between Exploration and Exploitation

Alina Ferecatu, Arnaud De Bruyn

To cite this article:

Alina Ferecatu, Arnaud De Bruyn (2022) Understanding Managers' Trade-Offs Between Exploration and Exploitation . Marketing Science 41(1):139-165. <https://doi.org/10.1287/mksc.2021.1304>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.



For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Understanding Managers' Trade-Offs Between Exploration and Exploitation

Alina Ferecatu,^a Arnaud De Bruyn^b

^aDepartment of Marketing Management, Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, Netherlands;

^bDepartment of Marketing, ESSEC Business School, 95000 Cergy, France

Contact: ferecatu@rsm.nl,  <https://orcid.org/0000-0003-0161-0487> (AF); debryun@essec.edu,  <https://orcid.org/0000-0002-4413-6167> (ADB)

Received: October 20, 2019

Revised: October 25, 2020; March 20, 2021

Accepted: April 3, 2021


Published Online in Articles in Advance:
October 21, 2021

<https://doi.org/10.1287/mksc.2021.1304>

Copyright: © 2021 The Author(s)

Abstract. Managers frequently explore new strategies, and exploit familiar ones, when making decisions on new product development, pricing, or advertising. Exploring for too long, or exploiting too soon, will generate inferior financial returns. Our research describes decision makers' exploration/exploitation trade-offs and their link to psychometric traits. We conduct an incentive-aligned study in which subjects play a multiarmed bandit experiment and evaluate how subjects balance exploration and exploitation, linked to psychometric traits. To formally describe exploration/exploitation trade-offs, we develop a behavioral model that captures latent dynamics in learning behavior. Subjects transition between three unobserved states—exploration, exploitation, and inertia—updating their beliefs about expected payoffs. Our analysis suggests that decision makers overexplore low-performing options, forgoing over 30% of potential revenue. They heavily rely on recent experiences. Risk-averse decision makers spend more time exploring. Maximizers are more sensitive to payoffs than satisficers. Our research builds the groundwork needed to devise remedial actions aimed at helping managers find an optimal balance between exploration and exploitation. One way to achieve this goal is by carefully designing the learning environment. In two additional studies, we analyze the evolution of exploration/exploitation trade-offs across different learning environments. Offering decision makers repeated opportunities to learn and increasing the planning horizon appears beneficial.

History: Avi Goldfarb served as the senior editor for this article.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "Marketing Science. Copyright © 2021 The Author(s). <https://doi.org/10.1287/mksc.2021.1304>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.

Funding: This work was supported by the Erasmus Research Institute of Management.

Supplemental Material: Data and the web appendix are available at <https://doi.org/10.1287/mksc.2021.1304>.

Keywords: managerial decision making • behavioral economics • exploration/exploitation trade-offs • multiarmed bandits • belief updating • satisficing

1. Introduction

In a wide range of business scenarios, managers need to strike a healthy balance between exploration and exploitation. Be it in marketing, research and development, pricing decisions, resource allocation, or new product development, to name but a few areas, assessing the expected value of the possible options requires experimentation (i.e., trial and error).

When adopting a strategy, managers need to search for alternative options while earning profits, and are thus faced with the dilemma of whether to explore or to exploit. Exploration behaviors involve search for novel options, experimentation, and innovation. Exploitation behaviors lean toward refinement, efficiency, selection, and implementation of known strategies

(March 1991). Either searching too much for the most appropriate option (consistent with exploration) or committing too fast to an option that may not be optimal (consistent with exploitation) could lead to reduced profitability. Understanding whether, why, and how managers systematically deviate from an optimal balance between exploration and exploitation is therefore an important research question.

Multiarmed bandits are a canonical representation of reinforcement learning problems. When choosing between several projects whose profitability is unknown, with the goal of maximizing earnings over time, managers are attempting to solve a complex version of a bandit problem. Solving the problem optimally requires decision makers to balance exploration

and exploitation, as they have to discover the profitability of the various options while also maximizing their earnings over time. Bandit problems therefore provide an interesting vehicle for studying decision makers' trade-offs between exploration and exploitation.

Multiarmed bandits are increasingly replacing A/B tests as tools to optimize the process of "earning while learning." In the recent academic literature, multiarmed bandit modeling has been incorporated into real-time website optimization (Hauser et al. 2014, 2009; Urban et al. 2014), online advertising (Schwartz et al. 2017, Baardman et al. 2019), and pricing problems (Misra et al. 2019). Liberali and Ferecatu (2019) propose a real-time optimization method in which a hidden Markov model is first used to infer consumers' position in the purchase funnel. This information is then integrated into a dynamic program that uses multiarmed bandits to dynamically match website design to consumers' information processing style, dependent on their position in the purchase funnel. Industry leaders also recommend experimental solutions based on multiarmed bandit problems. Google content experiments integrate bandit-like experimentation using Thompson sampling (Scott 2010).¹ The experimentation platform Optimizely offers a tool called the Stats Accelerator, described as a multiarmed bandit, to maximize click-through rates and conversions.² Amazon uses bandit experiments for real-time multivariate optimization of web content (Hill et al. 2017).

In this paper, we focus instead on a broader range of situations in which managers make decisions that can be conceptualized as multiarmed bandit problems but cannot be automated. Decision makers are expected to analyze the known facts, gather missing information, and formulate and evaluate alternative strategies before adopting a specific course of action. How managers solve such dynamic resource allocation problems is critical to the success of organizations. Our main contribution is therefore substantive. We aim to describe, quantify, and explain managerial exploration/exploitation trade-offs.

We conduct several studies in which subjects play a three-armed bandit experiment, set up as a prototypical managerial problem. Although subjects in our experiments are not managers, their decision making is informative of the trade-offs between exploration and exploitation expected when managers attempt to solve dynamic resource allocation problems in the field. Using controlled experiments rather than observational data allows us to understand decision makers' learning behavior, while eliminating other extraneous factors inherent in a complex business setting that might hinder our ability to understand the fundamental learning mechanisms at play. The bandit problem is

followed by a survey, in which we assess subjects' risk aversion, whether they used an analytical or intuitive style of decision making to tackle the task, and whether their general tendency is to maximize or to satisfice. We later use these psychometric traits as predictors of decision makers' learning tendencies.

First, to gauge the extent of the potential inefficiencies due to suboptimal learning, we compute the optimal exploration/exploitation strategy using the Gittins index (Gittins et al. 2011), a widely used policy for multiarmed bandit problems. We then compare subjects' search behavior to the optimal policy. We find that, on average, subjects forgo over 30% of the potential revenue, a finding that highlights the consequences of suboptimal behavior.

Second, to describe decision makers' learning tendencies potentially leading to suboptimal payoffs, we develop a behavioral model. Over the course of the bandit experiments, subjects do not only update their beliefs about the profitability of the options via sampling, but also change their sampling strategies over time, as they transition between exploration and exploitation. Subjects might also enter a state of inertia, where they minimize their cognitive efforts by repeating their previous choice without much consideration. We infer decision makers' unobserved transitions between exploration, exploitation, and inertia using an individual-level hidden Markov model. The specification is nonstationary and allows the outcomes of the bandit experiment to impact the transition probabilities between strategies. The belief-updating process follows the experience-weighted attraction (EWA) specification (Camerer and Ho 1999). We thus account for various consistent behaviors reported in the literature on decisions from experience (Erev and Haruvy 2015); these include discounting of previous payoffs or of previous experience with particular options, and how sensitivity to rewards affects the probability of choice.

Third, our paper builds the necessary blocks to assist managers optimally balance exploration and exploitation by predicting their idiosyncratic learning tendencies. Managers' psychometric traits can explain their learning tendencies and how they update their beliefs about the profitability of the various options. Using data from our main study (Study 1), we link subjects' inferred learning patterns to their attitudes toward risk, their analytical or intuitive decision-making style, and their tendency to maximize or satisfice. This allows us to anticipate their trade-offs between exploration and exploitation. In terms of the key results, we provide a rich understanding of how decision makers' learning behavior unfolds throughout the bandit experiment and document extensive heterogeneity in exploration/exploitation trade-offs. We find that outcomes that are disappointing relative to subjects' expectations affect how they transition

between exploration and exploitation. Our analysis shows that subjects are prone to forgetting earlier payoffs and to discounting their previous experience. Risk-averse decision makers are more likely to keep exploring options than those who are risk seeking. Maximizers are more sensitive to payoffs than satisficers. Our paper therefore diagnoses the suboptimal behavioral tendencies of decision makers. It paves the way to introducing a set of guidelines intended to help managers find an optimal balance between exploration and exploitation, possibly counterbalancing their natural tendencies.

One way to achieve this goal is by carefully designing the learning environment. In two additional studies (Studies 2 and 3), we investigate how changes in the learning environment, as reflected by the features of the bandit experiment, affect decision makers' exploration/exploitation trade-offs. We focus on one relevant dimension: the decision time frame. This dimension can be easily manipulated in a managerial setting. In Study 2, we allow subjects to play a multi-armed bandit experiment twice and document their "learning-to-learn" behavior over repeated experiments. In Study 3, we vary the planning horizon over which subjects learn. We show that increasing the planning horizon and offering decision makers the opportunity to repeatedly learn how to solve dynamic resource allocation problems can affect their learning behavior and bring them closer to the optimal path.

This paper is structured as follows. We introduce the relevant literature on managerial exploration/exploitation trade-offs in Section 2. We discuss our experimental design and describe the data from our bandit experiment in Section 3. In Section 4, we define the optimal sampling policy using the Gittins index, and we compare the behavior observed in our first bandit experiment to the optimal policy. In Section 5, we develop a behavioral model to capture exploration/exploitation trade-offs and their link with psychometric traits. We apply our model to the data from the bandit experiment and discuss the results in Section 6. In Web Appendix Section WA1, we show how decision makers' exploration/exploitation trade-offs are impacted by changes in the learning environment. In Section 7, we highlight the theoretical and substantive implications of our study and conclude by discussing what additional steps need to be taken to develop a prescriptive theory that can assist managers in their decision making.

2. Exploration/Exploitation Trade-Offs in Managerial Learning

We now review the literature relating the two building blocks of our theoretical development. We first discuss the literature on exploration/exploitation

trade-offs in bandit problems and elaborate on the likely impact of psychometric traits on learning behavior. We then link our study to the literature on managerial decision making.

2.1. Reinforcement Learning and Bandit Problems

In a multiarm bandit problem, a decision maker chooses repeatedly between several options, referred to as the "arms" of the bandit. The goal is to maximize the overall rewards over a defined period of time. Each arm that is chosen generates random rewards from a stationary distribution unknown to decision makers. The only way to discover which options may be most profitable is through sampling. Decision makers therefore face the classic exploration/exploitation dilemma, as they must balance learning about the profitability of each alternative with maximizing overall rewards over time.

Bandit-style problems are used to model a range of business- or economics-related decisions, and are found in various literatures, including operations research (Gans et al. 2007), organizational learning (March 1991, Posen and Levinthal 2011), and marketing (Meyer and Shi 1995, Lin et al. 2015, Shahrokh Tehranian and Ching 2019).

Several studies use an experimental paradigm to investigate how people tackle bandit problems. Most studies show that subjects use Bayesian updating to form expectations about the distributions of rewards, but also exhibit important and systematic deviations from a Bayesian-updating model. Gans et al. (2007) test two sets of models to represent how consumers choose between firms supplying products with different quality distributions. The two sets of models include heuristic approximations of normative models, and specifications that stem from the statistical learning literature. Less complex models provide a closer match to subjects' choice behavior. Meyer and Shi (1995) show that decision makers exhibit suboptimal behavior when tackling Bernoulli bandit problems, including a tendency to underexperiment with promising options and to overexperiment with ones that are less promising. The paper notes that subjects are not optimal learners, but they are forward looking, albeit over a limited planning horizon. Moreover, task characteristics such as the planning horizon and the success rates associated with the bandit's arms influence sampling behavior. Horowitz (1975) also documents decision makers' tendency to overexplore low-performing options, when success rates are low. Banks et al. (1997) manipulate the structure of the bandit problem such that the optimal solution is either myopic or forward looking, and find that subjects behave in line with these normative predictions. Anderson (2001) sets up more complex bandit tasks with normally distributed rewards and finds that,

by comparison with optimal behavior, subjects under-experiment. Risk aversion associated with diffuse priors on the distributions of rewards is the likely cause of this effect. In the cognitive psychology literature, Steyvers et al. (2009) use a modeling approach to test whether an optimal model better describes subjects' behavior against several heuristics and find that only 30% of subjects behave according to the optimal model. Several studies in cognitive neuroscience investigate human performance in bandit problems and how it relates to brain activity (Daw et al. 2006, Cohen et al. 2007, Ahn et al. 2014).

The studies described above compare decision makers' choices against models that involve an optimal balance between exploration and exploitation, but do not explicitly model subjects' exploration/exploitation trade-offs. The latter is the goal of our research.

Several papers use descriptive, model-free measures of behavior. For instance, Steyvers et al. (2009) defines the amount of exploration as the number of times a subject chooses an alternative with fewer successes and fewer failures than other alternatives. Gans et al. (2007) use the number of consecutive choices of the same supplier as a measure of the extent of exploitation. In the organization science and management literature, several studies define engaging in exploratory versus exploitative strategies as undertaking nonroutine versus routine tasks (Adler et al. 1999; Benner and Tushman 2003, 2002), or tasks with long-term versus short-term orientations (Tushman and O'Reilly 1996). A stream of literature in cognitive science (Roth and Erev 1998, Busemeyer and Stout 2002, Biele et al. 2009, Nevo and Erev 2012, Erev and Roth 2014) models the probability of moving from exploration to exploitation, albeit assuming different underlying dynamics in behavior than in the model we propose. This work inspired our model development. We compare our behavioral model to the above studies in Section 5.

Although relatively little experimental work has been done on how people solve bandit problems, many developments have proposed normative solutions to the problem. The canonical solution was introduced by Gittins and Jones (1979). Their proposed approach involves computing an index for each arm and choosing the arm with the highest index in every period. Gittins and Jones (1979) showed analytically that, assuming agents are risk neutral, this approach is optimal for dynamic problems stretching over an infinite time period, and that it maximizes the expected discounted overall rewards. For problems with finite horizons, the Gittins index is used as a heuristic.

2.2. Expected Impact of Psychometric Traits

2.2.1. Risk Aversion. Solving a bandit problem involves making decisions under uncertainty; thus, risk preferences of decision makers are likely to influence

how they learn. Many studies involving bandit problems focused on analyzing how subjects choose between a safe option that yields the same reward every time and a risky option yielding variable rewards. In this setup, exploration implies risk-seeking behavior, as it involves willingly sampling options with higher variability. In behavioral simulations based on popular reinforcement learning models, March (1996) shows that, with experience, subjects are predicted to select less risky options; thus, they learn to be risk averse. Denrell (2007, 2005) and Denrell and March (2001) attribute this phenomenon to the "hot stove" effect, a consequence of the inherent asymmetry between the impact of good versus bad outcomes. Arms yielding good outcomes are more likely to be re-sampled; thus, decision makers will gain a better idea of their profitability. Bad outcomes decrease the likelihood that an option will be chosen, and therefore a decision maker has fewer opportunities to discover its profitability. More risk-averse subjects underexplore options that (randomly) yield bad outcomes. These results were corroborated by studies linking neural correlates of reinforcement learning to experienced risk (Niv et al. 2012). Steyvers et al. (2009) test these effects using correlational analysis and do not find risk aversion to have an effect on experimental measures of exploration and exploitation.

Our setup involves choices between options with different expected rewards, but with similar variability in rewards. Here, risk aversion, as diminishing marginal sensitivity to rewards, is essential in learning, leading to more extensive exploration.

2.2.2. The Tendency to Maximize or Satisfice. The heuristics investigated in bandit experiments are consistent with a "satisficing" model (Gilboa and Pazgal 2001, Gans et al. 2007), in line with Simon (1959). Such heuristics assume that subjects define a target, or an aspiration level, and keep exploiting an option as soon as it exceeds that target. A satisficer's exploration/exploitation trade-offs are inherently different from the learning behaviors of a maximizer. Maximizers strive to find the option with the highest expected rewards (Schwartz et al. 2002), and exploit only when the option they believe will lead to optimal rewards has been identified. Therefore, we expect maximizers will explore more than satisficers.

2.2.3. Analytical vs. Intuitive Decision-Making Styles.

A bandit problem involves maximizing a criterion. It can be regarded as an expression of subjects' cognitive ability. Focusing specifically on information about payoffs might lead to better rewards than learning intuitively about the options (Toplak et al. 2011). Managers' decision-making style could influence their strategic choices, as well as their performance in the

learning task. Novak and Hoffman (2009) established that decision makers exhibit differences in thinking styles across tasks. We thus investigate how our subjects' analytical versus intuitive decision-making style impacts their learning tendencies when they are undertaking the experimental task.

2.3. Managerial Decision Making

Consumer-oriented learning models have relaxed the assumption that behavior is fully rational, but for managerial decision making, this is still a prevalent view, with a few notable exceptions (for a review, see Goldfarb et al. 2012). Using a structural model that embeds the cognitive hierarchy model of Camerer et al. (2004), Goldfarb and Yang (2009) show that, during the dot-com crash, managers' strategic thinking ability increased tech firms' chances of survival. The characteristics of managers affect firm performance and are key determinants of their ability. Goldfarb and Xiao (2011) show that managers who are better educated and have more experience are more likely to enter less competitive markets. Firms led by more able managers are more likely to stay in business and achieve higher revenues, provided that they survive.

In this study, we extend the literature on managerial decision making and focus on modeling exploration/exploitation trade-offs. We document how decision makers' psychological profiles and changes in the learning environment affect such trade-offs. These are necessary steps when attempting to find decision makers suited for various managerial tasks, or to nudge them toward the optimal path.

3. The Bandit Experiment

This section describes our experimental design, in which subjects tackle a bandit problem (or "bandit experiment").

We conduct a laboratory study (Study 1) where we analyze subjects' learning behavior in a tightly controlled environment, using a student sample. Although a student sample is not representative of a population of managers, the underlying psychological mechanisms we investigate here are similar. We expect the marginal impact of risk aversion on exploration/exploitation behavior to be similar across different populations, even if the distribution of risk aversion differs between our subjects and managers. Indeed, several studies use laboratory experiments in which subjects from a student or a general population undertake managerial tasks. Amaldoss et al. (2000) conducted two such laboratory experiments to understand how different types of strategic alliances are formed. Using student samples, Cui and Mallucci (2016) investigated how coordination of a marketing channel is affected when channel members' decision

making is impacted by fairness concerns. An additional benefit here is that the relatively homogeneous sample of students makes it easier to identify any individual differences in behavior, driven by psychometric traits. Therefore, we use the data gathered in Study 1 to validate our model describing exploration/exploitation trade-offs, and to document the impact of psychometric traits on learning behavior.

In addition to Study 1, we conduct two online experiments, with more diverse samples, to test whether changing specific features of the learning environment, such as changing the planning horizon and offering repeated opportunities to learn, impacts decision makers' learning behavior. These studies are presented in Web Appendix Section WA1.³

In this section, we introduce the experimental design used across the three studies. The specific manipulations in Studies 2 and 3 are discussed in Web Appendix Section WA1.

3.1. Experimental Design and Procedure

Students from a large European business school participated in our laboratory study. We recruited subjects by posting an announcement on the university's recruitment platform, inviting students registered with the subject pool to take part in an experiment on individual decision making, in exchange for monetary compensation.

We organized 15 laboratory sessions, with about six subjects per session. Students completed their tasks in separate soundproof cubicles, each with his or her own computer. On the screen in front of each subject was the first page of his or her task, and all other computer programs were disabled so that the subject could not use any tools during the study. Upon completion of the study, subjects' compensation was recorded and paid by bank transfer.

Subjects were asked to play a three-armed bandit experiment for 100 rounds. They were told to imagine that they were product managers for an online store, in charge of new product research and development. The three arms of the bandit problem represented three online banners advertising a product prototype they proposed to introduce in the market. Subjects were not given any information about consumers' interest in the product and had a 100-day period in which to evaluate the business potential of their product.

Every day, they could advertise their product using one of the three banner ads. At the end of the day, subjects were informed of the number of clicks made by prospective customers on their chosen banner. The number of clicks was used as a proxy for consumers' interest in the product. The following day, the product manager could select a different banner ad.

The product's potential would be judged on the total number of times prospective customers clicked on

the ads over the test period. The objective was to maximize the total number of clicks, hence demonstrating market interest. To do so, subjects needed to find the banner ad that would generate the most clicks per day, on average. Their payment was proportional to the cumulative number of clicks generated throughout the business task. This feature ensured that the experiment was incentive aligned.

We showed subjects three buttons, labeled banners A, B, and C, corresponding to the three arms of the bandit. The three alternatives were shuffled at the start of the experiment, so no participant was able to know a priori which arm would yield the highest expected reward. Before the main task, subjects were given a 10-round trial period to familiarize themselves with the setting, after which the buttons corresponding to the banner ads were reshuffled for the main task.

We linked each button to a random number generator that generated rewards between 0 and 100. The rewards followed normal distributions with expected values of 65, 50, and 35, and a standard deviation (SD) of 15 for all the arms. Throughout our analysis, we label arms 1, 2, and 3 those with expected values of 65, 50, and 35 clicks respectively. Therefore, arm 1 is high performing, leading to the highest expected rewards, whereas arms 2 and 3 are low performing, leading to lower expected rewards on average. Subjects were informed that the rewards vary between 0 and 100 for all three banner ads, and that the three banner ads have similar and constant variation in rewards. We stressed in the instructions that the distributions behind the options remained constant throughout the exercise. Subjects were not informed about the shape of the distribution.

Prior to the experiment, we drew five sets of experimental stimuli from the distributions above and randomly assigned a set of draws to each subject. Given that our study took place in a physical laboratory across several days, this minimized the chance that subjects would have learned the optimal strategy from previous participants. We ensured that the five sets of draws led to similar optimal policies, to be able to compare subjects' learning behavior. Moreover, we ensured that all rewards were in the gains domain, as losses might lead to different patterns of behavior (Tversky and Kahneman 1992, Erev et al. 2008).

To increase the ecological validity of our study, we used a design that was based on a choice between three normally distributed arms, rather than using a one-armed or two-armed Bernoulli bandit, typically used in bandit experiments (Meyer and Shi 1995, Gans et al. 2007, Steyvers et al. 2009). We designed the scenario to portray a business situation in which exploration/exploitation trade-offs are salient.

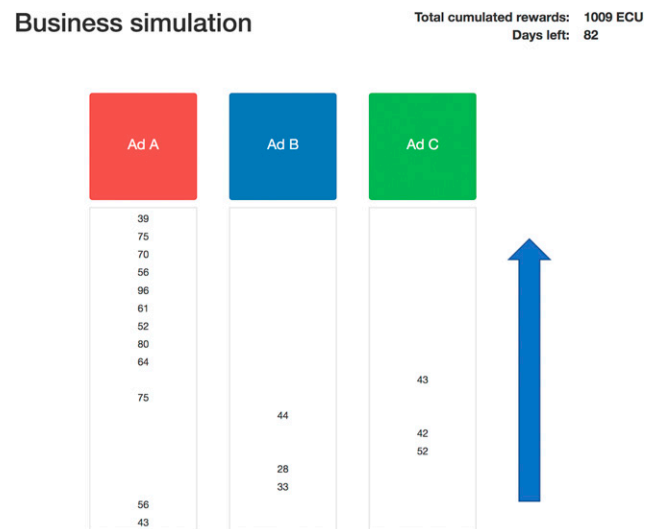
We chose the parametrizations of the distributions to ensure one of the arms would be a rather clear winner. We let participants play for 100 rounds to ensure

that the exploration and exploitation stages were sufficiently long. We expected significant variation in behavior, necessary for the parametric identification of our behavioral model.⁴

We expected subjects to first click all three buttons to gather information about the expected payoffs for each alternative (exploration), and then to identify with increasing certainty the button with the highest expected reward and commit to that option (exploitation); this is similar to the strategy followed by the subject depicted in Figure 1.

In the second stage of the experiment, subjects were asked to complete a psychometric questionnaire. To elicit their risk preferences, we used the “bomb test,” validated by Crosetto and Filippin (2013). The incentive-aligned task elicits subjects' constant relative risk aversion (CRRA) coefficient.⁵ To elicit subjects' maximizing or satisficing tendencies, we used a short version of the scale developed by Schwartz et al. (2002). The short scale consists of six items and was validated by Nenkov et al. (2008). We used a 20-item scale developed by Novak and Hoffman (2009) to determine subjects' situation-specific thinking style. This scale identifies whether subjects use a more analytical or intuitive approach when tackling a task. This scale includes two 10-item subscales, a “need for cognition” scale and a “faith in intuition” scale.⁶ We report summary statistics and scale reliability measures in Web Appendix Section WA2.

Figure 1. Screenshot of the Experimental Platform, with Choices and Rewards for One Subject



Notes. The figure shows the experimental interface (partial screen capture). Subjects were asked to maximize the overall rewards, over 100 rounds. As they chose between options, subjects received information on the payoffs from their chosen option and their cumulative payoffs. This subject sampled all options for the first 9 rounds (exploration process), and then focused on ad A for the next 9 rounds (exploitation process). We have added the arrow to indicate the sequence of results. ECU, experimental currency units. One click by prospective customers is worth 1 ECU.

3.2. Descriptive Statistics and Data Patterns for Study 1

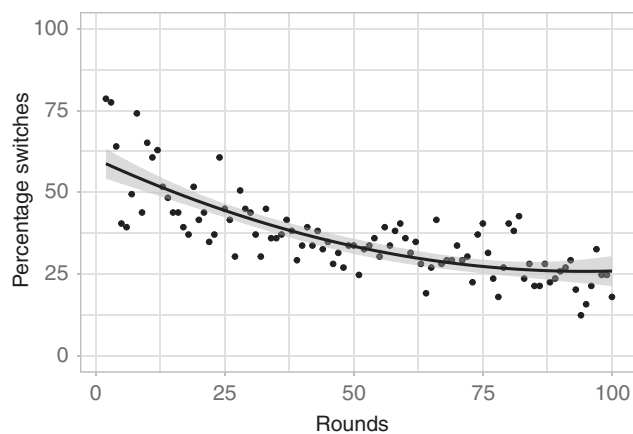
Eighty-nine students participated in Study 1 ($M_{age} = 21$, 28% male). The experiment took an average of 10 minutes, lasting at least 4.5 minutes and no more than 25 minutes. All students received a payment of EUR 1 for showing up, plus their cumulative rewards from the risk aversion task and the bandit problem. The average reward for the experimental task was EUR 4.93 (SD = EUR 0.22), at an exchange rate of 1,500 clicks to EUR 1. The average reward for the risk aversion task was EUR 0.68 (SD = EUR 0.78). The total average reward was EUR 5.62 (SD = EUR 0.84).⁷

We now document the data patterns describing how subjects chose between the options. As our goal is to understand behavior, we focus on the patterns that are illustrative of the underlying learning process (Shmueli 2010). During the first 10 rounds of the bandit experiment, subjects switched between options from one round to the next 59.2% of the time, whereas in the last 10 rounds, they did so only 22.1% of the time (see Figure 2). This suggests that, early on, subjects sampled different arms, presumably in an attempt to identify which banner ad to use. Toward the last rounds of the bandit experiment, subjects appear increasingly confident in their choice of the best arm, choosing it consistently. Two subjects never sampled all arms.

Subjects repeated their previous choices 63.1% of the time. This suggests significant state dependence in choice behavior, which we specifically model in Section 5.

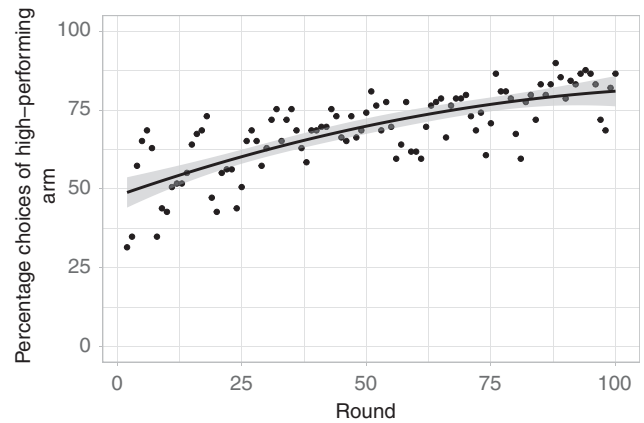
Whereas subjects selected the high-performing arm, leading to the highest expected reward about 33% of the time during the first few rounds (equal to chance), after 40 rounds, the hit rate quickly increased to above 60% (see Figure 3). In the second part of the bandit

Figure 2. Percentage of Subjects Switching to a Different Arm in Each Round



Notes. Subjects sampled different arms to a higher extent early on. Choices become more consistent in the later rounds of the bandit experiment, with subjects repeatedly choosing the same option.

Figure 3. Percentage of Subjects Choosing the High-Performing Arm 1 in Each Round



Note. The percentage increases over rounds, but remains below 100% even after 100 rounds, showing that there is room for improvement in subjects' learning strategies.

experiment, the hit rate stagnated at around 70%. The fact that the hit rate remained stable and was below 100% suggests that a significant proportion of subjects followed suboptimal learning paths.

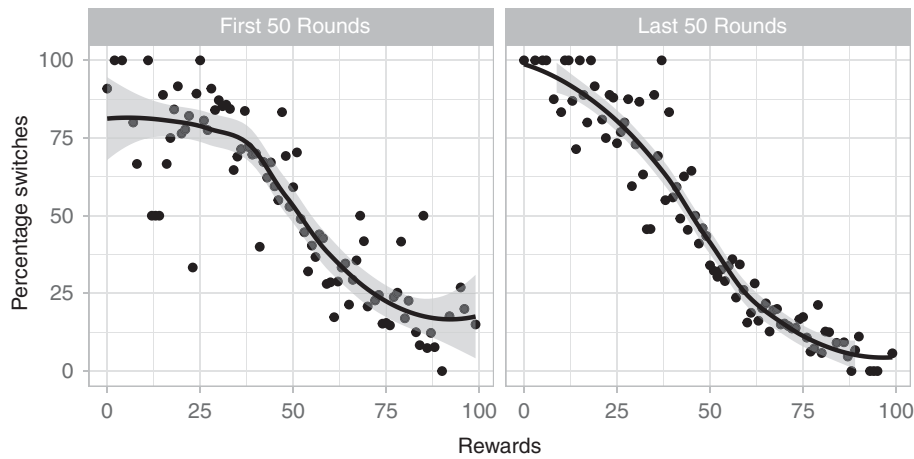
Figure 4 shows that the lower the reward received in the current round, the more likely subjects were to choose a different option in the next round. This effect appears even stronger in the last part compared with the first part of the experiment. Presumably, subjects in exploitation mode who receive disappointing rewards tend to switch to a different arm. The logistic regression coefficient of lagged rewards on the probability of switching options is -0.057 ($p < 0.01$) over the first 50 rounds, and -0.063 ($p < 0.01$) over the last 50 rounds. This suggests that low and high rewards impact in different ways the learning patterns of decision makers: disappointing results with one arm seem to induce subjects to explore other arms more.

At the aggregate level, the correlation between subjects' overall rewards and the total number of switches between arms is strongly negative at -0.74 ($p < 0.01$). This suggests that many subjects might overexplore, switching between arms too often instead of exploiting the high-performing arm 1.

The descriptive evidence presented here also shows that subjects paid attention to the task and strove to identify the best option. Subjects spent a decreasing amount of time per round as they played the bandit experiment, from an average of 1.9 seconds per round in the first 10 rounds to 0.9 seconds in the last 10 rounds. Figure 5 depicts the choices of the first three subjects and the time spent on each choice.

In 22.4% of the rounds, subjects spent less than half a second on their choice, and in 92.4% of those rounds, subjects reinforced their last choice. This again suggests that choices are state dependent. It would be

Figure 4. Percentage of Subjects Switching to a Different Arm Across the Distribution of Previous Rewards



Notes. The lower the previous reward, the more likely subjects are to switch to a different arm. This effect appears stronger in the last 50 rounds of the bandit experiment, compared with the first 50 rounds, probably because disappointing rewards have a stronger effect in the exploitation phase than in the exploration one.

virtually impossible for subjects to update their expectations about the profitability of the options in such a short amount of time.

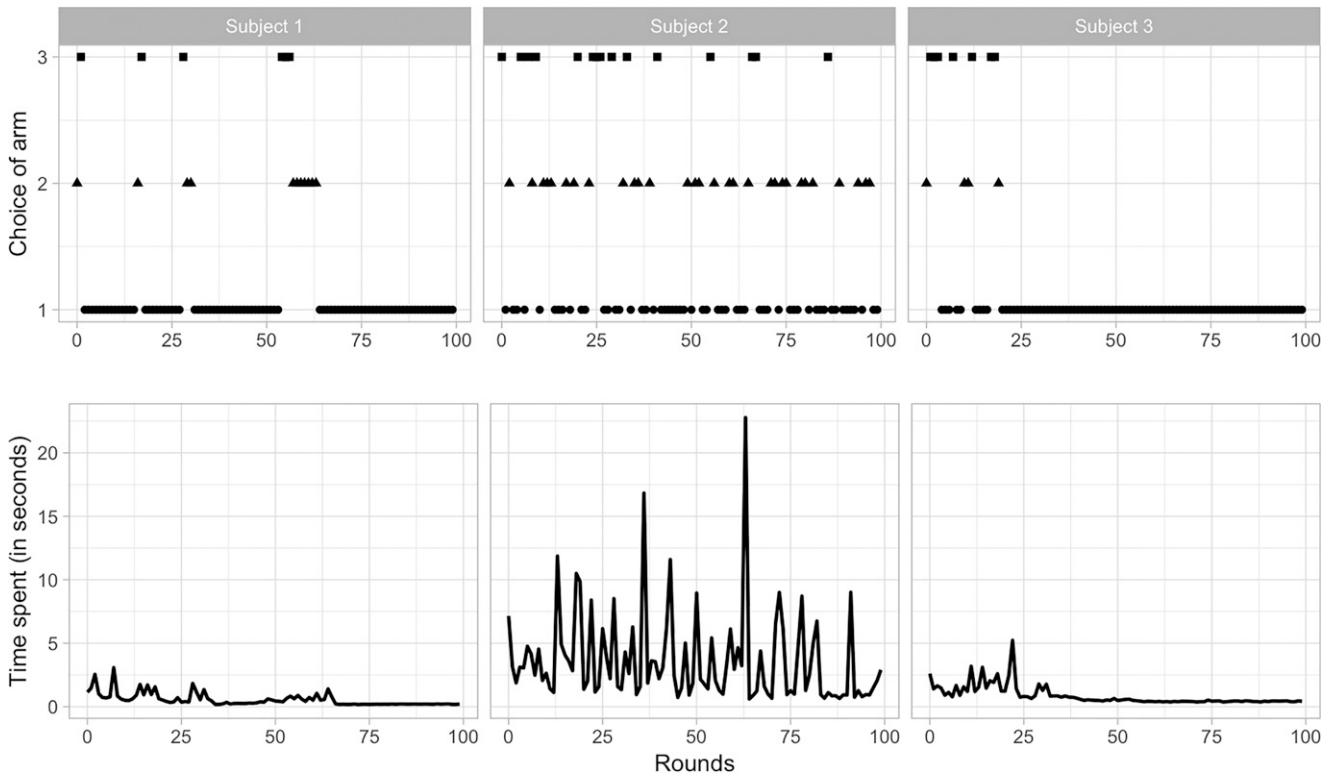
The rewards subjects received in the previous round correlate with the time they spent on deciding between arms. The higher the previous reward, the less time they spent on their decision in the following round ($cor(Lagged\ rewards, Time\ spent) = -0.12, p < 0.01$).

Subjects spent more time on the choice task in the round following disappointing rewards.

4. The Optimal Path

To highlight the importance of understanding managerial learning, we quantify the extent of potential payoffs subjects forgo, as they attempt to balance exploration and exploitation. To do this, we compute

Figure 5. Choice of Arm and Time Spent on the Decision for Subjects 1, 2, and 3, Exhibiting Different Learning Patterns



Note. Subjects tend to spend more time to decide on rounds in which they switch between arms.

the optimal path they *could* have pursued instead. We briefly discuss the optimal path computation for the multiarmed bandit experiment played by subjects, then compare subjects' overall rewards and their sampling patterns to the optimal learning policy.⁸

4.1. Optimal Path Computation

In a multiarmed bandit problem, a decision maker samples several options (or “arms”) multiple times. The arms are indexed by $j = 1, \dots, J$. Each arm delivers random, normally distributed rewards X_j , with unknown mean μ_j , variance σ_j^2 , and precision $A_j = 1/\sigma_j^2$. In our experimental task, the mean rewards of the three arms are unknown to subjects and are set at $\mu_j = \{65, 50, 35\}$. The standard deviation is assumed to be known and equal for all of the three arms, set at $\sigma^X = 15$. The purpose is to maximize the total discounted rewards over a time frame T . We set the discount factor to $a = 0.99$, a common value in many bandit experiments with humans. The discount factor reflects the weight companies put on future outcomes and can be application-specific.⁹

Instead of solving a J -dimensional dynamic program to optimally find the path that maximizes expected discounted rewards, Gittins and Jones (1979) proposed an optimal solution using an index policy. An arm-specific index I_j^t is computed at every round t . Gittins and Jones (1979) and Gittins et al. (2011) showed that for an infinite-horizon problem, at each round t , it is optimal to choose the arm with the highest index. Such a policy would maximize the expected discounted rewards, assuming agents are risk neutral. For a finite time horizon, the Gittins index is a

heuristic resulting in a near-optimal policy for multiarmed bandits with normally distributed rewards (Lattimore 2016).¹⁰

Figure 6 plots the evolution of the Gittins indices, the posterior mean rewards, and the value of exploration for one set of draws used in our experiment, assuming an agent follows the optimal path.

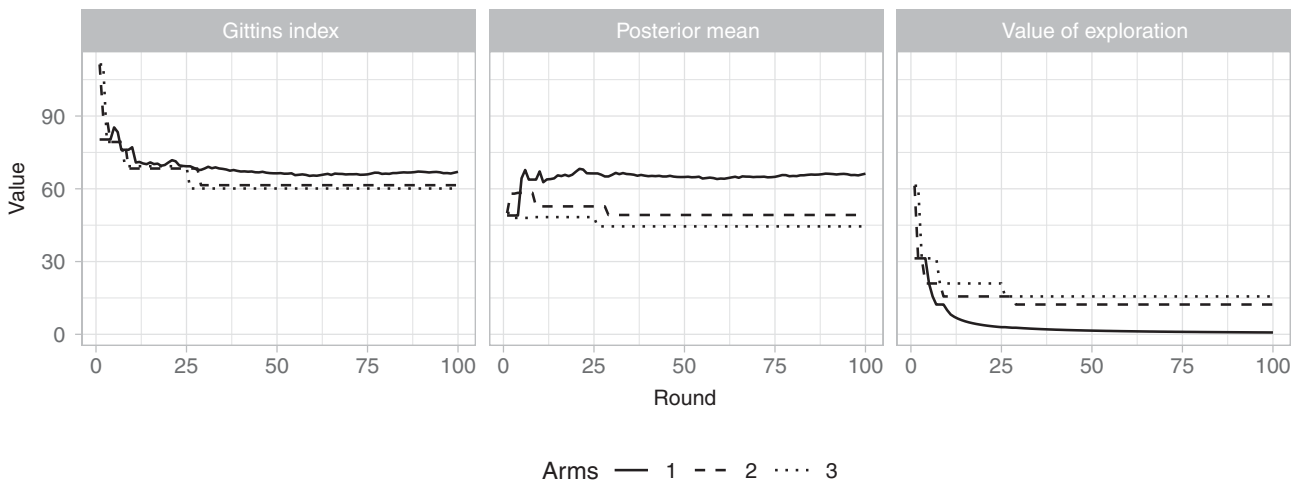
The value of exploration is the difference between the Gittins index computed for each arm and the posterior mean rewards. In the first few rounds, uncertainty is high and there is value in exploring. The Gittins indices plotted in Figure 6 have higher values in earlier rounds compared with later rounds. As uncertainty decreases, the value of exploring different arms decreases.

4.2. Optimal Path vs. Actual Behavior in Study 1

We now discuss how subjects' learning patterns differ systematically from optimal exploration/exploitation trade-offs, and whether and how their overall rewards differ from optimal reward levels, using the data gathered in Study 1. Note that we withheld certain pieces of information that subjects would have needed to compute Gittins indices, and subjects were not allowed to use any tools throughout the experiment. Therefore, subjects could not compute indices and follow the optimal path strictly. Our main goal here is thus to quantify the extent of potential payoffs they forgo, payoffs that would be obtained by following an optimal learning policy.

4.2.1. Comparison of Overall Rewards. Assuming decision makers have perfect information about the

Figure 6. The Evolution of the Gittins Indices, the Posterior Mean Rewards, and the Value of Exploration When Agents Follow the Optimal Path



Notes. An agent following the optimal path starts by sampling each arm at least once, in random order. Throughout the bandit experiment, the agent samples low-performing arms 2 and 3 seven times. Starting round 26, the agent repeatedly samples the high-performing arm 1 through to round 100. The value of exploration is high initially and decreases as the agent gathers more information about the distributions of rewards.

distribution of rewards behind each arm, they would systematically select the high-performing arm. At the other extreme, a decision maker not engaging in learning would sample the options randomly throughout the bandit experiment. We compute overall rewards per subject under the above benchmarks using the experimental draws and average these overall rewards across subjects. We use the range between the perfect information and the random benchmarks to quantify how subjects in our experiment deviate from the optimal policy. Table 1 reports the results of this analysis.

Following the optimal policy leads to overall rewards of 6,379 clicks (SD = 156 clicks) on average, with returns close to 88.3% of the perfect-information benchmark. Subjects in our experiment earn overall rewards of 5,890 clicks (SD = 336 clicks) on average. Consider the overall rewards under random sampling as the lower bound and the overall rewards under optimal learning as the upper bound of potential revenues in the experimental task. In Table 1, we show that the relative efficiency of the experimental rewards is 63.2%. Therefore, on average, subjects in Study 1 forgo 36.8% of potential revenues.

There is considerable variation in the overall rewards earned by subjects. Figure 7 shows significant overlap between the distributions of experimental rewards and those of the optimal and the random benchmarks. In fact, for 3.4% of subjects, overall rewards are above the optimal benchmark. At the lower end, for 2.3% of subjects, overall rewards acquired in the experiment are below the random benchmark.

4.2.2. Comparison of Sampling Behavior. For each subject, we compute the percentage of low-performing arms 2 and 3 sampled throughout the experimental task to give us a descriptive measure of the *actual* extent of exploration. Subjects sample low-performing options 31.9% of the time on average (standard error (SE) = 0.017). We compare this statistic to the percentage of low-performing arms sampled under optimal behavior to give us a descriptive measure of the *optimal* extent of exploration. Under optimal behavior, a decision maker samples low-performing arms 8.3% (SE = 0.004) of the time on average. In line with Horowitz

(1975), our results show that the majority of subjects oversample low-performing options. The results are in line with those of Steyvers et al. (2009), who found that the optimal model best explained behavior for only 30% of the subjects. We noted above that under a Gittins policy, agents are assumed to be risk neutral. Relaxing this assumption by using a Whittle index (Whittle 1988, Lin et al. 2015, Shahrokhi Tehrani and Ching 2019) to compute the optimal path would lead to more extensive sampling of low-performing arms.

To describe decision makers' trade-offs between exploration and exploitation in more detail, we propose a behavioral model in the next section.

5. The Behavioral Model

5.1. Model Overview

We use a dynamic modeling approach to characterize the reinforcement learning behavior of decision makers. In line with our expectations, we find evidence to support the view that subjects engage in exploration/exploitation trade-offs, as they attempt to reach the overarching goal of maximizing their overall rewards over the given time frame. The descriptive evidence presented in Section 3.2 suggests extensive state dependence in subjects' choices, as they likely enter a state of inertia.

In exploration, subjects' goal is to learn the profitability of the various options. The sampling process is highly probabilistic and only loosely related to the expected payoff of the arms; subjects need to sample all options repeatedly and construct beliefs about the rewards distributions associated with each arm.

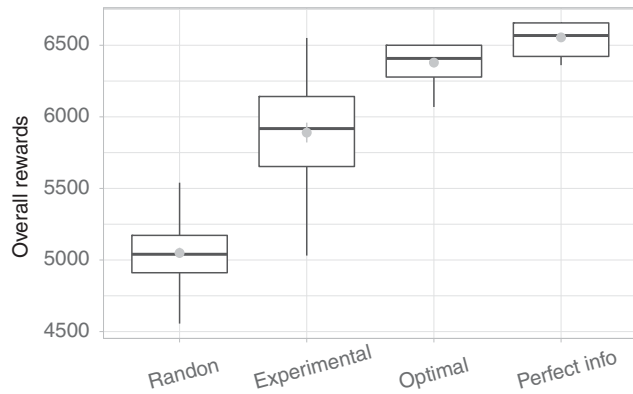
In exploitation, the objective is to repeatedly use the option the subject believes will lead to the highest expected payoff at every round. This strategy is therefore more deterministic than an exploration strategy, although choices still include a random component. If a subject mostly chooses the high-performing arm but occasionally samples the second best-performing arm, they will be estimated in exploitation, because choices are not fully deterministic and subjects can make mistakes. At the extreme, when two options are estimated to yield similar expected rewards, switching between

Table 1. Benchmark Comparisons: Difference in Average Overall Rewards with Experimental vs. Optimal Learning

	Average overall rewards (in clicks)	Improvement over random choice (%)	Efficiency vs. perfect information (%)	Efficiency vs. optimal learning (%)
Random	5,049	0	0	0
Experimental	5,890	16.7	55.8	63.2
Optimal learning	6,379	26.3	88.3	100
Perfect information	6,555	29.8	100	—

Notes. Improvement over random choice is computed as $(\overline{OR}_{Policy_i} - \overline{OR}_{Random}) / \overline{OR}_{Random}$. Efficiency versus perfect information is computed as $(\overline{OR}_{Policy_i} - \overline{OR}_{Random}) / (\overline{OR}_{Perfect\ Info} - \overline{OR}_{Random})$. Efficiency versus optimal learning is computed as $(\overline{OR}_{Policy_i} - \overline{OR}_{Random}) / (\overline{OR}_{Optimal} - \overline{OR}_{Random})$. The term \overline{OR} stands for average overall rewards.

Figure 7. The Distribution of Overall Rewards Earned by Subjects Throughout the Experimental Task, Plotted Against the Random, Optimal, and Perfect Information Benchmarks



Notes. The box-and-whisker plots represent the full distributions of the measures. The means (grey dots) and confidence bounds are overlaid.

them at random, while avoiding the other options, can still be considered as exploitation.

The choice process conditional on using an exploration or an exploitation strategy is partly driven by the expected rewards. Several heuristics are likely to influence how subjects update their beliefs about the profitability of the options, and thus their choice behavior. Subjects might have various degrees of sensitivity to rewards, they might discount previous payoffs, or they might underestimate the amount of information accumulated about each option.

In a state of inertia, subjects do not update their beliefs about expected rewards, but simply reinforce their last choice of arm. Several studies made on decisions from experience show that subjects' choices reveal a strong positive correlation, which implies inertia (Erev and Haruvy 2005). Strong state dependence is particularly apparent in cases where the decision in previous rounds required sustained cognitive effort and the cost of updating beliefs is higher than the expected benefit (Erev and Haruvy 2015).

The exploitation and the inertia states involve conceptually distinct psychological mechanisms. However, as we observe only choices and rewards, the choices in exploitation and in inertia are close to observationally equivalent. We discuss here the data patterns identifying whether subjects' choices are likely driven by exploitation or by inertia.

The model infers that subjects who repeatedly choose the arm with the highest expected payoff are initially in exploitation and switch to inertia after several rounds of repeating the same choice. This is in line with the patterns in the data observed in our experiment, particularly in terms of time spent on choices (see Figure 5), with subjects spending a decreasing amount of time on each choice as they repeatedly reinforce the previous choice of arm.

Descriptive evidence presented in Section 3 supports an inertia state. In 22.4% of rounds, subjects in our study spend less than half a second on decisions. In 92.4% of those rounds, they reinforce their last choice or arm. This suggests that subjects reinforce their last choice without updating their beliefs about the expected payoffs. Therefore, inertia can explain a sampling strategy where, after reaching exploitation, a subject keeps choosing an arm despite increasing evidence that the expected rewards are inferior to those of the arms not chosen. It is unlikely to rationalize such behavior under the exploitation state.

Even when in exploitation, subjects can switch between arms and occasionally choose lower-performing arms, because choices are stochastic and driven by the expected payoffs of each arm. Subjects cannot switch between arms when in inertia, because here choices are deterministic and independent of expected payoffs. To illustrate the data patterns that differ between exploitation and inertia, consider the following: if, after initial exploration, a subject mostly chooses the high-performing arm over a few dozen rounds, but occasionally the subject samples the second-best-performing arm, this subject is most likely in exploitation. If, after initial exploration, a subject chooses only the high-performing arm over a few dozen rounds, the model may infer that the subject is initially in exploitation, and after several rounds of exploitation, the subject is likely to switch to the inertia state.

State dependence emerges as a likely component necessary to explain learning behavior, both theoretically and based on the data patterns presented in Section 3. In the next section, we compare various model specifications and focus on in-sample fit measures to determine whether the inertia state is instrumental in understanding behavior. We show this is indeed the case. We investigate the predictive performance of the models and discuss under what conditions a model including the inertia state is useful to predict exploration/exploitation trade-offs.

The three states of exploration, exploitation, and inertia, which we also refer to as "sampling strategies," are unobserved. We assume that, throughout the bandit experiment, subjects transition between these sampling strategies. We observe only which options subjects sampled and the associated rewards. Our modeling approach uses a hidden Markov model to uncover subjects' latent transitions between the three sampling strategies, and their choices conditional on the strategy implemented in each round.

Outcomes of the bandit experiment can trigger transitions between sampling strategies. Descriptive evidence in Figure 4 indicates that the lower the rewards obtained, the more likely subjects are to choose a different option. This suggests that subjects are more likely to change their strategy after receiving disappointing

payoffs (Ansari et al. 2012). Nevo and Erev (2012) show that individuals are more likely to exit the inertia state when the outcomes are surprising enough to engender a change of strategy, and outcomes that are disappointingly low surprise subjects. In line with the “hot stove” effect (Denrell and March 2001; Denrell 2007, 2005), subjects shy away from options that randomly generate low outcomes and do not learn effectively about their payoff distributions. Arms with better outcomes have a higher probability of being chosen, and thus increase learning about the distribution of their rewards. Meyer and Shi (1995) find that subjects are more likely to switch away from an arm after a disappointing outcome, to a higher degree than would be predicted by a Bayesian learning model.

Although our behavioral model is flexible in allowing for any transition patterns, we expect disappointing outcomes to affect changes of state in different ways, depending on the sampling strategy subjects have been using. For instance, we expect disappointing outcomes to increase the likelihood that subjects stay in exploration and to decrease the likelihood of them moving from either exploration or exploitation to inertia, because they pay more attention to rewards. We also expect a disappointing outcome to decrease subjects' probability to remain in a state of inertia.

Our proposed behavioral model has two components: (1) subjects' state transitions between sampling strategies and (2) subjects' choice behavior conditional on the sampling strategy used. We model decision makers' behavior at the individual level to account for individual heterogeneity versus latent dynamics in sampling strategies and to evaluate the impact of psychometric traits on learning.

5.2. Transitions Between Sampling Strategies

We assume $K = 3$ latent states of exploration ($k = 1$), exploitation ($k = 2$), and inertia ($k = 3$), reflecting subjects' sampling strategies. Subjects transition between these sampling strategies over time. The strategy used by decision maker i at round t , K_{it} , evolves over time following a first-order Markov decision process with nonstationary and heterogeneous transition probabilities.

The nonstationary probability of transitioning from state k at round $t - 1$ to state k' at round t follows a multinomial logit model:

$$q_{ikk'(t-1)t} = \begin{cases} \frac{\exp(\beta_{0ikk'} + \beta_{1kk'}D_{t-1})}{1 + \sum_{m=1}^{K-1} \exp(\beta_{0ikm} + \beta_{1km}D_{t-1})}, & \text{for } k = 1, \dots, K, k' = 1, \dots, K-1; \\ 1 & \\ \frac{1}{1 + \sum_{m=1}^{K-1} \exp(\beta_{0ikm} + \beta_{1km}D_{t-1})}, & \text{for } k = 1, \dots, K, k' = K. \end{cases} \quad (1)$$

Parameters $\beta_{0ikk'}$ represent individual and state-specific propensities to transition between sampling strategies. The inertia state ($k = 3$) is used as the baseline, and the utility of transitioning from state k into inertia is set to zero for identification purposes. State-specific parameters $\beta_{1kk'}$ capture the impact of the time-varying covariates D_{t-1} on the transition probabilities. Note that $\beta_{1kk'}$ are not individual specific; thus, we implicitly assume that the heterogeneity in transition behavior is captured by the baseline propensities.

The time-varying covariate D_{t-1} classifies the outcome of the bandit experiment at round $t - 1$ as either disappointing or encouraging relative to the subject's expectations. Expectations EV_{t-1} are operationalized as the running average of each subject's payoff up to and including the previous round $t - 2$.

The variable D_{t-1} is defined as

$$D_{t-1} = I_{(-\infty, 0)}(EV_{t-1} - X_{t-1}) \quad (2)$$

The term $I_{(-\infty, 0)}(a)$ is an indicator function that classifies an outcome as disappointing (equal to one) when $a \in (-\infty, 0)$ and encouraging (equal to zero) otherwise.¹¹ This specification is in line with the work of Erev and Haruvy (2015), who show that previous losses relative to players' expectations impact their strategic choices. In Roth and Erev (1998), choice reinforcement is shown to depend on the difference between payoffs and an updated reference point. This is consistent with the work of Nevo and Erev (2012), who show that when players have access to feedback on the outcomes of earlier options, their future behavior is impacted by payoffs they have forfeited by not taking those options. The result is reinforced in the work of Ansari et al. (2012), who show that in a multi-player game, using a low-performing option in the previous round could trigger a transition between the rules used by subjects to learn about the profitability of the options.

As in the paper by Nevo and Erev (2012), we assume that all subjects start the bandit experiment using an exploration strategy; thus, the initial probabilities of exploration, exploitation, and inertia are given by $\pi_{0k} = [1, 0, 0]$, respectively.

5.3. Belief Updating

Descriptively, a bandit problem as a general class of reinforcement learning can be modeled using a combination of cumulative and averaged reinforcement learning specifications.

In cumulative reinforcement learning, options have reinforcement utilities that increment cumulatively with current rewards and impact the choice likelihoods. In averaged reinforcement learning, payoffs

are averaged rather than cumulated over time, such that reinforcement utilities are bounded by the payoff distributions.

We use the EWA model (Camerer and Ho 1999), which allows for a mix of the two types of reinforcement learning. Ho et al. (2006) note the appropriateness of the EWA specification for modeling how managers learn over time. Rapoport and Amaldoss (2000) use the EWA model to study whether managers making investment decisions engage in iterative elimination of strongly dominated strategies, or mixed strategies. The specification can converge to reinforcement learning that is either entirely averaged or entirely cumulative, if the data provide evidence to support it.

For subject i , at time t , an option j has a numerical attraction $A_{ij}(t)$. Each option's attraction is updated with the experience at time t as

$$A_{ij}(t) = \frac{\phi_i N_i(t-1) A_{ij}(t-1) + X_{ij}(t)}{N_i(t)}, \quad (3)$$

$$N_i(t) = \rho_i N(t-1) + 1,$$

where ϕ_i decays past attractions, ρ_i decays past experience with the arm, and $N(t)$ can be interpreted as the number of "observation equivalents" of past experience.

We assume initial attractions $A_{ij}(0)$ to be linked to prior beliefs. In the experiment, we inform subjects that rewards are limited between 0 and 100; therefore, a reasonable prior for $A_{ij}(0)$ would be 50. Prior experience reflects the strength of belief in the prior distribution. We set $N_i(0)$ to one, and assume that the prior belief had a strength of one "experience equivalent."¹²

When $N_i(0) = 1/(1 - \rho_i)$ and $\phi_i = \rho_i$, the EWA model is reduced to an averaged reinforcement learning model (Sarin and Vahid 1999, Busemeyer and Stout 2002). When $N_i(0) = 1$ and $\rho_i = 0$, the model reduces to the cumulative reinforcement learning model (Roth and Erev 1995).

The attractions of the arms are computed as a mix of cumulative performance and average performance. With $1 > \phi_i > \rho_i$, attractions fall between running averages and the running total. When two arms perform equally well, the arm selected more frequently is rated somewhere between equally as good as and twice as good as an arm that has been drawn half as much. This shows that subjects focus on arms that have been vetted before, valuing their experience with the arms. The reverse is true when $1 > \rho_i > \phi_i$. When two arms perform equally well, the more frequently sampled arm is rated between half as good and equally as good as the arm that has been drawn half as much. Subjects focus on exploring arms not vetted sufficiently. This property of the model is relevant in our bandit problem, particularly for early choices.

5.4. Choice Probabilities Conditional on Subjects' Beliefs

The probability that subject i chooses arm j at round t , p_{ijt} , is a function of the attractions at round $t - 1$ and follows a logit specification:

$$p_{ijt} = \frac{\exp(\lambda_i A_{ij}(t-1))}{\sum_{l=1}^J \exp(\lambda_i A_{il}(t-1))}. \quad (4)$$

Conditional choice probabilities are equivalent to logit transformations of the attractions of options, similar to a multinomial logit model widely used in marketing (Ho et al. 2006, Cui and Mallucci 2016). The positive-definite parameter λ_i captures subjects' sensitivity to attractions, and therefore to rewards obtained. As λ_i increases, the arm with the highest attraction is more likely to be chosen. The heterogeneity in λ_i can be viewed as differences in subjects' sensitivity to rewards. Web Appendix Section WA5 further describes the evolution of attractions and their impact on choice probabilities in the EWA model.

5.5. Choice Conditional on Sampling Strategies

5.5.1. Choice Under Exploration. When using an exploration strategy, subjects learn about the profitability of the options by sampling them. Following Equation (4), this implies that we expect the reward sensitivity parameter $\lambda^{Explore}$ to be close to zero, as beliefs about the profitability of each option do not strongly impact choice. We specify $\lambda^{Explore}$ at the aggregate level.

5.5.2. Choice Under Exploitation. When in exploitation, subjects focus on maximizing the expected payoffs. The attractions of each alternative impact the choice probabilities, following Equation (4). The individual-level sensitivity parameter, $\lambda_i^{Exploit}$, is expected to be higher than $\lambda^{Explore}$, but not infinite (to allow some randomness even in exploitation). For identification purposes, we impose an order constraint on the two sensitivity parameters, such that $\lambda^{Explore} < \lambda_i^{Exploit}$.

5.5.3. Choice Under Inertia. In inertia, subjects reinforce the previous choice. The choice probabilities are specified as

$$p_{ijt}^{Inertia} = I(J_t = J_{t-1}), \quad (5)$$

where $I(\cdot)$ is an indicator function, taking a value of one when the arm chosen is the same as the one chosen in the previous round and zero otherwise.

Because subjects do not focus on realized payoffs and do not update their beliefs about the arms' profitability, the inertia state allows them to keep sampling low-performing arms, even when there is increasing evidence that these arms are indeed low performing. Once subjects exit inertia, they update their beliefs about the profitability of the options by

discounting the entire history of payoffs and follow an exploration or an exploitation strategy when making their choices.

In our setup, choice probabilities conditional on sampling strategies come from different distributions. Whereas choice probabilities in exploration and exploitation come from the same distribution, but with different parameters, choice probabilities in the inertia state have a degenerate distribution. Therefore, our model is effectively a hidden Markov mixture of experts model (Ansari et al. 2012).

5.6. Heterogeneity and the Impact of Psychometric Traits

We allow for heterogeneity in the belief-updating process and in the propensities to transition between sampling strategies, linked to decision makers psychometric traits. This ensures that we properly disentangle learning dynamics from variation between subjects. All individual-level parameters, $\{\beta_{0ikk'}\}$, $\lambda_i^{exploit}$, ρ_{it} , and ϕ_{it} , are gathered in $vec(\delta_i)$, and vary following a multivariate normal distribution, with mean μ_{δ_i} , and a full covariance matrix Σ . We transform the elements in $vec(\delta_i)$ such that each component varies over the real line, while certain individual-level parameters are bounded.¹³ We break down the dependency between the state-specific group-level parameters and the individual-level component to ensure proper identification of the model parameters at both layers of the hierarchy. We sample individual-level parameters using a noncentered reparametrization (Betancourt and Girolami 2015):

$$\begin{aligned}\delta_i &= \mu_{\delta_i} + L\Sigma\xi_{it}, \\ \mu_{\delta_i} &= \Gamma z_{it},\end{aligned}\quad (6)$$

where $L\Sigma$ is the Cholesky factor of the covariance matrix Σ , and $vec(\xi_i) \sim N(0,1)$. This effectively shifts the correlation between the data and the parameters to the hyperparameters. We gather vectors ξ_{it} , which are uncorrelated, in the Ξ matrix. We decompose the covariance matrix into a location and a scale prior, such that $\Sigma = \text{diag}(\tau)\Omega\text{diag}(\tau)$. The term $\text{diag}(\tau)$ is a diagonal matrix of scale parameters and Ω is the correlation matrix. The term z_{it} is a vector of the mean-centered psychometric variables and includes an intercept. The matrix of coefficients Γ reflects the impact of the psychometric variables on the behavioral parameters.¹⁴

Conditional on the chosen sampling strategy, the likelihood that a decision maker i will choose arm j at round t is given by

$$\begin{aligned}\mathcal{L}(\mathbf{B}_0, \beta_1, \rho, \phi, \lambda^{Explore}, \lambda^{Exploit}, \Gamma, \Xi, \tau, \Omega \mid \text{data}) \\ = \sum_{k_{t1}=1}^K \sum_{k_{t2}=1}^K \dots \sum_{k_{tT}=1}^K \{P(K_{t1} = k_{t1}) \\ \prod_{t=2}^T P(K_{it} = k_{it} \mid K_{i(t-1)} = k_{i(t-1)}) \times \prod_{t=1}^T P(Y_{ijt} = y_{ijt} \mid K_{it} = k_{it})\end{aligned}\quad (7)$$

We use the forward-filtering technique (Murphy and Bach 2012) and a hierarchical Bayesian approach involving the Hamiltonian Monte Carlo (HMC) sampling method to estimate the model parameters (Gelman et al. 2013).¹⁵

5.7. Comparison with Previous Learning Models

Our modeling approach is distinct from previous learning models that inspired it in several ways, which we highlight below.

Our conceptualization of underlying and unobserved sampling strategies is similar to the inertia, sampling and weighing (I-SAW) model proposed by Nevo and Erev (2012), where subjects exhibit three response modes: exploration, exploitation, and inertia. We differ in how we model subjects' transitions between response modes and how they choose between options under exploration and exploitation.

In the I-SAW model, choices in exploration and exploitation are rule based. Decision makers choose randomly (uniform distribution) under exploration, and choose with certainty the alternative with the highest estimated subjective value in exploitation. We allow decision makers to deviate even in exploitation from the assumption of subjective value maximization, and decision makers have varying degrees of sensitivity to rewards. This sensitivity is imposed to be higher in exploitation than in exploration, such that choices are more random in exploration.

Choices under inertia are similarly conceptualized in both models, and the probability of inertia decreases with the difference between expected and observed outcomes, in line with Rescorla and Wagner (1972). But in the I-SAW model, both positive and negative outcomes relative to expectations can impact the inertia state probability. This is because subjects receive complete feedback about obtained and forgone payoffs. Our assumption is that only disappointing outcomes trigger changes in sampling strategies, because feedback is limited to obtained outcomes. Positive outcomes relative to expectations confirm subjects' beliefs about the profitability of the chosen arm and are unlikely to trigger a change in behavior.

An important assumption that sets our model apart from other reinforcement learning models, such as the exploitative sampler, the contingent sampler (Biele et al. 2009), or the I-SAW model, is that sampling strategies follow a first order Markov decision process. Thus, the probability of selecting sampling strategy k at round t depends on the sampling strategy selected at round $t - 1$.

We implement a myopic model. In our specification, decision makers do not solve a dynamic program by integrating the future value of information, or the "exploration bonus" in their choices, as they would in a forward-looking model. Evidence of forward-looking

behavior was documented by Meyer and Shi (1995), who found that a two-period look-ahead model fits observed choice patterns in bandit problems better than an optimal or a myopic model. Yang et al. (2015) and Lin et al. (2015) show that including a forward-looking component in dynamic discrete choice models improves predictions of consumers' preferences. Shahrokh Tehrani and Ching (2019) compare a myopic benchmark to several forward-looking models to understand and predict consumer learning. They show that a heuristic approach based on the value of perfect information captures consumers' behavior effectively and is computationally efficient. Our assumption is that decision makers intuitively understand that there is value in exploring, and we account for this goal by introducing the exploration state.

Gans et al. (2007) recommend using an exponential smoothing model, nested in a logit specification with a sensitivity parameter, to describe subjects' behavior in bandit problems. We propose a generalized version of this model, increasing our ability to describe decision makers' trade-offs between exploration and exploitation. In addition to accounting for how subjects forget early rewards, we account for how subjects disregard previous experience with an arm. The sensitivity parameter differs across exploration and exploitation modes. Gans et al. (2007) acknowledge that subjects spend very little time on choices as they learn to play the game, but do not integrate this behavior in their modeling approach. We account for this state dependence by introducing the inertia state.

To summarize, our behavioral model, hereafter labeled the EEI/EWA model (i.e., exploration–exploitation–inertia model with experience-weighted attractions), is a novel integration of several components. First, Markov dynamics explain how subjects choose between three sampling strategies, exploration, exploitation, and inertia. Second, choices under exploration and exploitation are governed by the EWA model, whereas choices under inertia are driven by the immediate past. Third, learning dynamics are individual specific and informed by subjects' psychometric traits.

6. Quantifying Exploration/Exploitation Trade-Offs and the Impact of Psychometrics

We apply the behavioral model specified in the previous section to our experimental data from Study 1 to describe subjects' exploration/exploitation trade-offs.

We start by discussing subjects' behavior as inferred by our model. This highlights the complexity of the model and the relationships between its various moving parts. Last, we compare our model to various

benchmarks to show how these moving parts are informative of decision makers' learning behavior.

6.1. Parameter Estimates

We first discuss decision makers' transitions between sampling strategies, followed by their belief-updating process and choices conditional on sampling strategies. We then describe the impact of psychometric traits on learning. Table 2 reports all group-level parameter estimates and their 95% highest density intervals (HDIs).¹⁶

6.1.1. Transitions Between Sampling Strategies. We capture dynamics between latent states by allowing subjects to transition between sampling strategies across rounds. Table 3 shows the posterior distributions of the transition propensities for the average decision maker, computed based on the behavioral parameters reported in Table 2.

To highlight the impact of experimental outcomes on state transitions, we report in Table 3 two sets of transition matrices, one following encouraging outcomes, and one following disappointing outcomes. The large diagonal elements of the transition matrices suggest that subjects have a tendency to repeat the previous sampling strategy.

If we compare the fourth and the seventh columns of Table 3, we observe that subjects are less likely to remain in inertia after a disappointing outcome ($q_{Inertia-Inertia}^D = 0.477$) than after an encouraging one ($q_{Inertia-Inertia}^E = 0.751$); disappointing results prompt decision makers to reevaluate their strategy and

Table 2. Group-Level Estimates of Behavioral Parameters

Parameter	Mean	95% HDI	
Baseline transition parameters			
$\hat{\beta}_{0,11}$	1.605	0.260	2.885
$\hat{\beta}_{0,12}$	0.409	-1.292	1.948
$\hat{\beta}_{0,21}$	-2.779	-4.097	-1.355
$\hat{\beta}_{0,22}$	1.224	-0.130	2.620
$\hat{\beta}_{0,31}$	-2.517	-4.362	-0.839
$\hat{\beta}_{0,32}$	-1.383	-2.971	0.231
Impact of disappointing outcomes			
$\hat{\beta}_{1,11}$	1.373	-0.279	3.024
$\hat{\beta}_{1,12}$	0.479	-1.375	2.169
$\hat{\beta}_{1,21}$	0.012	-2.059	2.003
$\hat{\beta}_{1,22}$	2.688	1.060	4.109
$\hat{\beta}_{1,31}$	1.209	-0.662	2.461
$\hat{\beta}_{1,32}$	1.194	0.008	2.271
Belief-updating parameters			
$\hat{\lambda}^{Explore}$	0.082	0.058	0.105
$\hat{\lambda}^{Exploit}$	0.187	0.095	0.335
$\hat{\phi}$	0.221	0.141	0.310
$\hat{\rho}$	0.693	0.640	0.740

Notes. The table shows means and 95% HDIs of the (transformed) behavioral parameters. The inertia state 3 is used as the baseline.

Table 3. Means and 95% HDIs for the Posterior State Transitions, Following Disappointing or Encouraging Outcomes

From\to	State transitions					
	Following disappointing outcomes			Following encouraging outcomes		
	Exploration	Exploitation	Inertia	Exploration	Exploitation	Inertia
Exploration	0.851 [0.478, 0.855]	0.105 [0.0339, 0.142]	0.043 [0.002, 0.488]	0.665 [0.504, 0.691]	0.201 [0.107, 0.271]	0.134 [0.039, 0.389]
Exploitation	0.001 [0.001, 0.002]	0.979 [0.717, 0.997]	0.019 [0.001, 0.283]	0.014 [0.009, 0.017]	0.762 [0.464, 0.916]	0.224 [0.067, 0.528]
Inertia	0.129 [0.006, 0.276]	0.394 [0.049, 0.668]	0.477 [0.054, 0.945]	0.061 [0.012, 0.161]	0.188 [0.048, 0.468]	0.751 [0.371, 0.939]

presumably invest more cognitive effort in the task. Similarly, decision makers are more likely to enter an inertia state after a good result ($q_{Exploration-Inertia}^E = 0.134$, $q_{Exploitation-Inertia}^E = 0.224$) than after a disappointing one ($q_{Exploration-Inertia}^D = 0.043$, $q_{Exploitation-Inertia}^D = 0.019$).

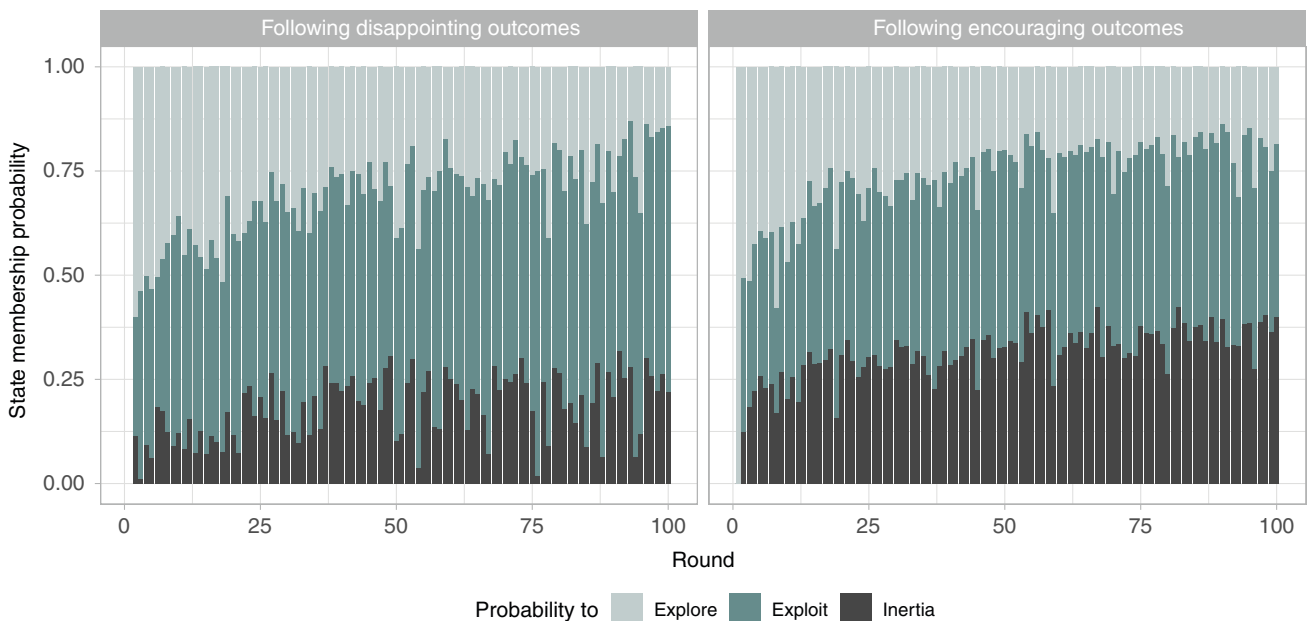
For every subject, at every round of the bandit experiment, we compute the probabilities of using each sampling strategy, and we plot the results in Figure 8. There are substantial dynamics in sampling strategies. Throughout the bandit experiment, as expected, the likelihood of exploring decreases, whereas the likelihoods of exploitation or inertia increase. After disappointing outcomes, subjects are much less likely to use an inertia strategy.

6.1.2. Choice Conditional on Sampling Strategies.

Figure 9 presents the predicted probability densities of choosing each arm, when using a sampling strategy of exploration, exploitation, or inertia. In exploration, there is a large overlap between the probabilities of choosing arms 1, 2, or 3. When in exploitation, subjects are likely to choose the high-performing arm 1 and highly unlikely to choose the lowest-performing arm 3. In an inertia state, subjects are likely to mechanically repeat their choice of arm 1, although at times they may enter inertia while focusing on low-performing arms 2 and 3.

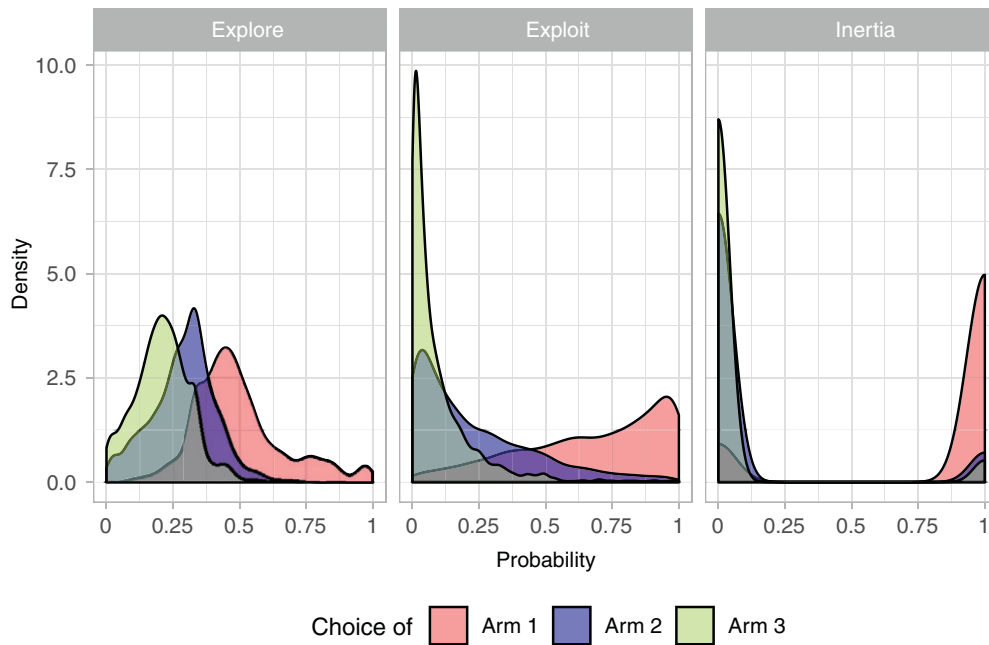
Subjects' choice of arm in exploitation is dependent on the belief-updating process, governed by the EWA model. With reference to Table 2, the average decay

Figure 8. Probabilities of Sampling Strategies Across Rounds, Following Disappointing or Encouraging Outcomes



Notes. Subjects start in exploration and are more likely to stay in exploration than to move from exploration to exploitation or to inertia (see also Table 3). This leads to more choices in exploration early on. Once in exploitation or inertia, subjects are unlikely to switch to exploration. We therefore observe a larger probability to explore early on, and a lower probability to explore in later rounds of the bandit experiment. This mirrors the exploration/exploitation trade-offs expected in bandit problems, and highlighted in the description of our experimental data (see Section 3.2).

Figure 9. Choice Probability Densities Conditional on Sampling Strategy



Notes. Choices in exploration appear more random compared with choices in exploitation. The choice probabilities conditional on being in inertia are one for the previously chosen arm and zero otherwise. Subjects are likely to repeat their choice of arm 1 when in inertia, although at times they may focus on the low-performing arms 2 and 3. Therefore, the conditional choice probabilities of these arms have some density at one, but much less so compared with the conditional probability of choosing arm 1.

parameter $\bar{\phi}$ (mean = 0.221, 95% HDI = [0.141, 0.310]) is lower than $\bar{\rho}$ (mean = 0.693, 95% HDI = [0.640, 0.740]), implying that attractions grow more slowly than each arm's average payoffs. This suggests that attractions are highly dependent on recent outcomes and on the extent of experience with an arm, and explains subjects switching behavior between states.

6.1.3. Heterogeneity in Exploration/Exploitation Trade-Offs. There is significant heterogeneity in state transition patterns across individuals, and in how subjects update their beliefs about the profitability of the three options.¹⁷

Figure 10 displays the subject-specific state probabilities across rounds against subjects' actual choices. The model is able to capture subjects' sampling strategies efficiently. Subject 3 strikes a good balance between exploration, exploitation, and inertia. The subject explores extensively in the first 20 rounds, then enters a state of inertia. Subject 4 is estimated to mostly explore, focusing too much on the least effective arm 3. Subject 46 engages in exploration in the first few rounds, rapidly decides to exploit arm 1, and then keeps sampling that arm for the rest of the bandit experiment.

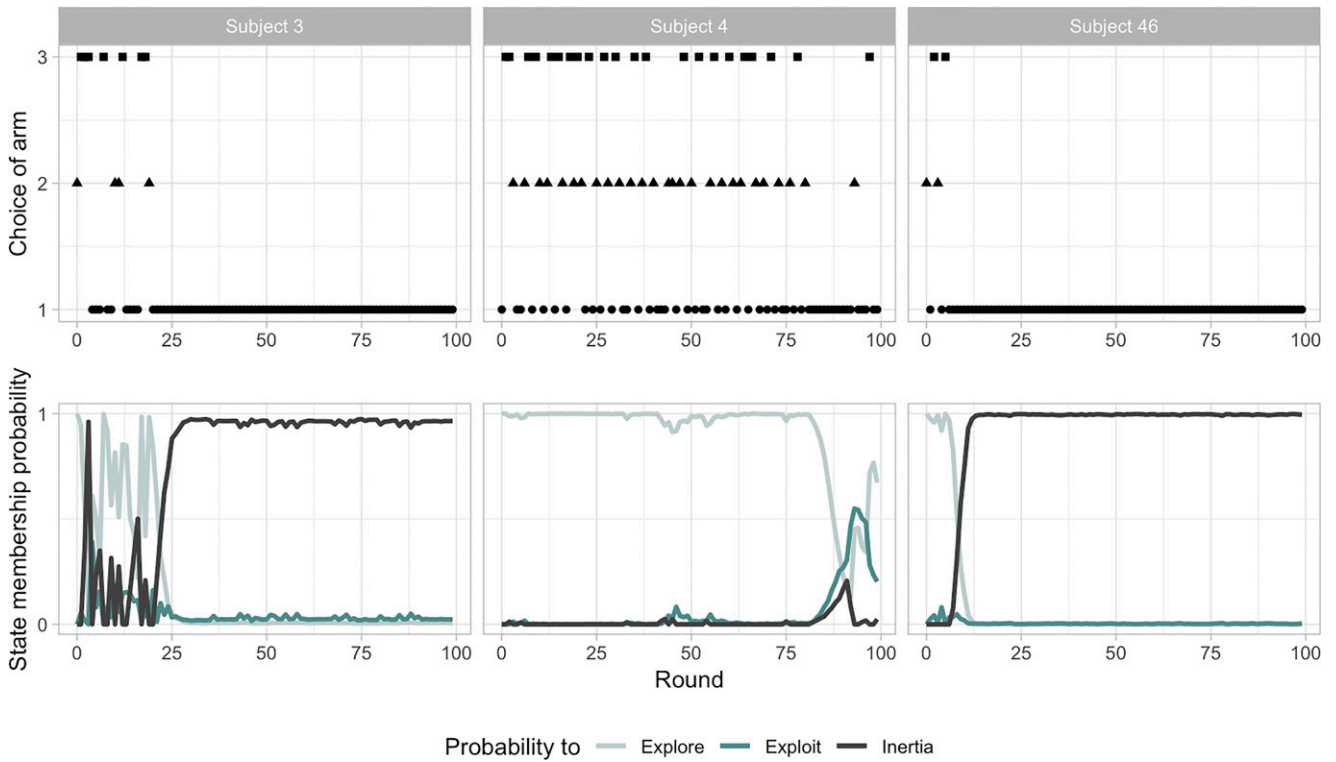
6.1.4. The impact of Psychometric Traits. Figure 11 plots the probability densities of the parameter estimates showing the impact of psychometric traits on learning behavior.¹⁸

More risk-seeking decision makers are less likely to stay in exploration ($\delta_{CRRA, \beta_{0i,11}} = -1.094$, 95% HDI = [-1.987, -0.274]).¹⁹ This is in line with our conjecture in Section 2. Unsurprisingly, maximizers are more sensitive to the attractions of options ($\delta_{Maximiz, \lambda_i^{Exploit}} = 0.220$, 95% HDI = [0.009, 0.546]), as they are trying to find the option leading to highest payoff, rather than the option merely reaching a minimum threshold of acceptability. Subjects who stated they have used a more analytical decision-making (ADM) style to solve the bandit problem tend to discount more previous experience ($\delta_{ADM, \rho_i} = -0.264$, 95% HDI = [-0.506, -0.023]). Because the average decay rate of past experience is much higher than the average decay rate of past attractions, a heavier rate of decay for past experience would bring its level of discounting closer to that of past attractions.

A manager cannot alter the psychological profile of a task leader, but can choose the leaders to be assigned to various tasks. Our analysis enables managers to link a task leader's psychological profile to their exploration/exploitation trade-offs and anticipate their learning tendencies.²⁰

6.2. Comparing Our Model to Relevant Benchmarks

The main goal of this study is to understand and explain exploration/exploitation trade-offs. Given this, we built a model informed by theory that captures the

Figure 10. Subject-Level Choice of Arm and State Membership Probabilities Across Rounds

Notes. We plot choice behavior (top) and the evolution of the state membership probabilities for three subjects (bottom), with various sampling patterns. Arm 1 leads to the highest expected rewards, followed by arms 2 and 3.

most relevant features of the data to thoroughly describe learning behavior. This can come at the expense of predictive ability (Shmueli 2010). First, building more complex models leads to a decrease in bias, but can increase variance (Hastie et al. 2016). Second, fitting many features of the model in-sample allows us to thoroughly explain behavior, but it can hurt a model's flexibility to predict new data out-of-sample. In this section, we start by discussing in-sample fit measures to check whether our behavioral model is most informative of the main features of the data when compared with several relevant benchmarks. We then assess the robustness of the proposed model in terms of its predictive validity. We end by discussing the ability of our proposed model and various benchmarks to capture patterns of behavior out-of-sample.

Our behavioral model has two main components: (a) the hidden Markov model that assumes probabilistic transitions between exploration, exploitation, and inertia (EEI) and (b) a belief-updating component (EWA). We first compare our full EEI/EWA model to two nested models and highlight the necessity of both model components. We call this set of nested models the “main components” benchmarks:

- *EEI model*. This is a hidden Markov model that excludes the belief-updating component. Choices under

exploration are random ($\lambda^{Explore} = 0$), and choices under exploitation are quasi-deterministic ($\lambda^{Exploit} = 1$). Choices in inertia remain as specified in Equation (5). The expected payoffs are computed as arm-specific running averages of rewards earned up to and including the previous round ($\rho = \phi = 1$).

- *EWA model*. The expected payoffs are computed using the EWA model, following Equation (3). Choice probabilities include the sensitivity parameter λ_i . The model is static; it does not include the hidden Markov component.

The second set of benchmarks, labeled “key feature” benchmarks, highlights the interest of having more fine-grained adjustments to the model, nested within the main components discussed above:

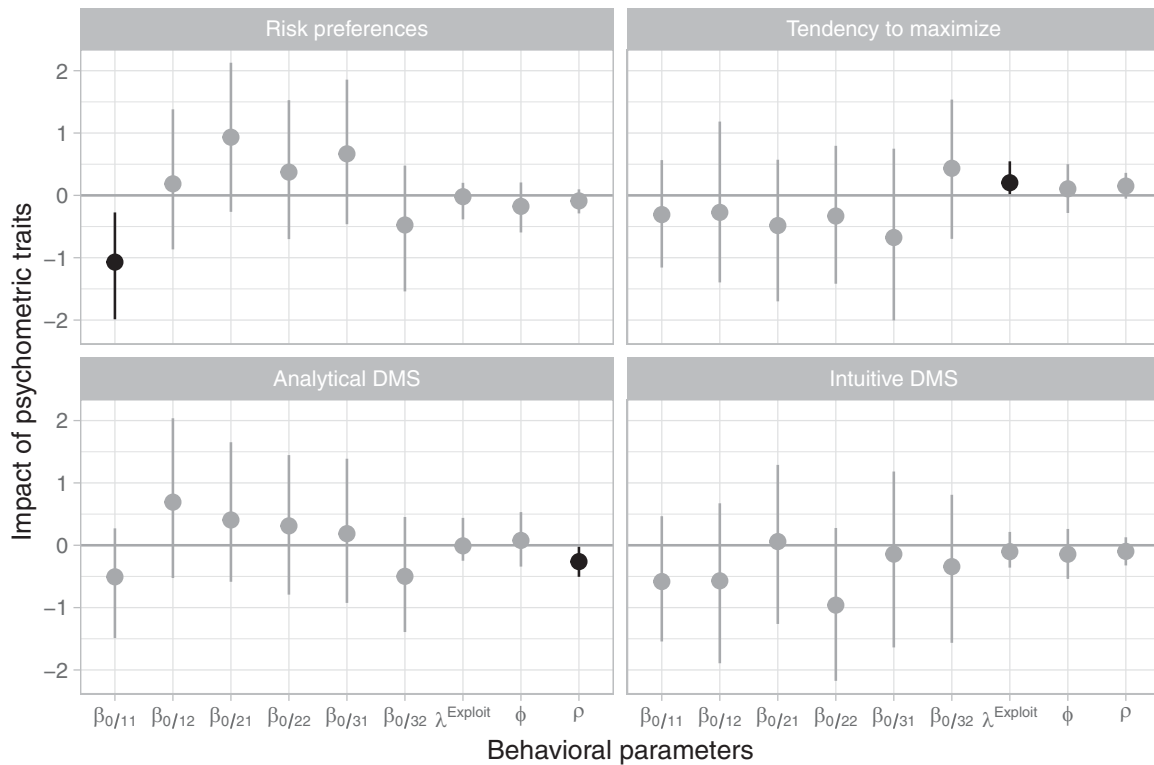
- *statEEI/EWA*. This is a stationary hidden Markov model in which the individual-specific transition probabilities between sampling strategies are constant throughout the bandit experiment ($\beta_{1kk'}$ are set to zero).

- *EE/EWA*. This is another nested model that excludes the inertia state; subjects only move between exploration and exploitation.

- *EEI/EWA no PsychVars*. This model excludes subjects' psychometric traits. In this model, Equation (6) includes only an intercept.

To assess the fit of our model against the above benchmarks, we compute in-sample likelihood- and

Figure 11. Means and 95% HDIs of the Parameters Estimating the Impact of Psychometric Traits on the (Untransformed) Behavioral Parameters



Notes. Parameter estimates for which the 95% HDIs do not include zero are in black. DMS, decision-making style.

non-likelihood-based measures and discuss our model’s ability to recover key statistics of the data. We then discuss the models’ ability to predict choices out-of-sample using a temporal forecasting and a cross-sectional validation approach.

6.2.1. In-Sample Fit. All models were estimated using the same Bayesian framework and the same prior distributions as in the proposed EEI/EWA model. To assess each model’s in-sample fit, we report in Table 4 the log-predictive density and the Watanabe–Akaike information criterion (WAIC; Gelman et al. 2013) as likelihood-based measures of fit. WAICs account for model fit while correcting for model complexity and adjust for the number of parameters. We also report mean squared errors (MSEs) and hit rates as likelihood-free measures of fit. For each individual at each round and each iteration (after convergence) of the HMC sampler, we first infer the underlying sampling strategy a subject follows, drawn based on the filtered state probabilities predicted by the model. We then infer a choice of arm conditional on the realized sampling strategy, using the predicted choice probabilities conditional on state. We compute the squared error as the square of one minus the predicted probability of the chosen arm. We average this measure across

rounds and across individuals to obtain an MSE. Therefore, in our bandit experiment with multinomial choices, MSE as the difference between predicted probabilities and observed choices is a measure of the confidence of each model in its predictions (Ansari et al. 2012, Ascarza and Hardie 2013). The hit rate is the average percentage of correctly predicted choices across individuals and iterations of the HMC sampler.

In Table 4, with respect to the main components benchmarks, in-sample fit measures support the dynamic EEI/EWA model that allows subjects to transition between sampling strategies over the static EWA model. The EEI model, which excludes the belief-updating component, fits the data best when looking at likelihood-free criteria (hit rate at 74.3%, MSE at 0.171), but its WAIC is significantly worse compared with all other models.

With respect to the fey features benchmarks, the in-sample fit of the EEI/EWA model is superior to the in-sample fit of the statEEI/EWA and EE/EWA models. This supports the conjecture that our dynamic model that includes nonstationary transition probabilities between sampling strategies and an inertia state is best suited to explain learning behavior here.

The EEI/EWA model that excludes psychometric variables fits the data similarly to the proposed model

Table 4. Model Comparison—In-Sample Fit

Model	Log-pred. density	WAIC	MSE	Hit rates	MAE
EI/EWA (proposed model)	−5,780	10,886	0.189	68.8%	0.092
EWA	−5,807	11,483	0.220	63.2%	0.118
EI	−6,075	11,745	0.171	74.3%	0.078
stat EI/EWA	−5,805	10,940	0.190	68.5%	0.092
EI/EWA	−5,726	11,209	0.200	67.5%	0.100
EI/EWA no PsychVars	−5,772	10,869	0.189	68.9%	0.091

in terms of likelihood-free criteria. The WAIC of the model without psychometric traits is lower than the WAIC of the proposed model. This is expected, as the measure penalizes the proposed model for the additional 36 parameters in the heterogeneity specification, included to study the impact of psychometric traits on exploration/exploitation trade-offs. Estimating the additional parameters can impact the variance of the group-level parameters, which in turn negatively affects WAICs (Gelman et al. 2014). It should also be noted that EI/EWA model without psychometrics retains the ability to capture idiosyncratic behaviors at the individual level (through the individual-level intercepts in the Bayesian specification). Consequently, the fact that it achieves similar in-sample fit comes at no surprise. Stripped of psychometric variables, however, it cannot capture the underlying—and systematic—sources of such individual differences. Because our main goal is to understand and explain learning behavior, we retain the EI/EWA model.

We present posterior predictive checks as additional measures of in-sample fit to further describe how various models recover key features of the data. A key statistic in our bandit experiment is the subject-level percentage of switches between arms, as a proxy of the extent of learning subjects engage in. We test how well the EI/EWA model and the various benchmarks recover this statistic. We compute the absolute error as the absolute difference between the predicted and observed percentages of switches between arms for each subject and at each iteration of the HMC sampler. Table 4 reports the mean absolute error (MAE) in predicting the percentage switches across subjects. To further understand how different learning behaviors are recovered by the models, we split the sample into deciles based on the observed percentage of switches between arms. This will break down the performance of models for subjects who engage in limited sampling (5.4% switches on average for subjects in Decile 1) versus those who engage in extensive sampling of the options (85.2% of switches on average for subjects in Decile 10). Figure 12 plots per decile the predicted versus observed mean percentage of switches.

First, the EI/EWA model accommodates well the behavior when there is little to a moderate amount of switching between arms. However, the model

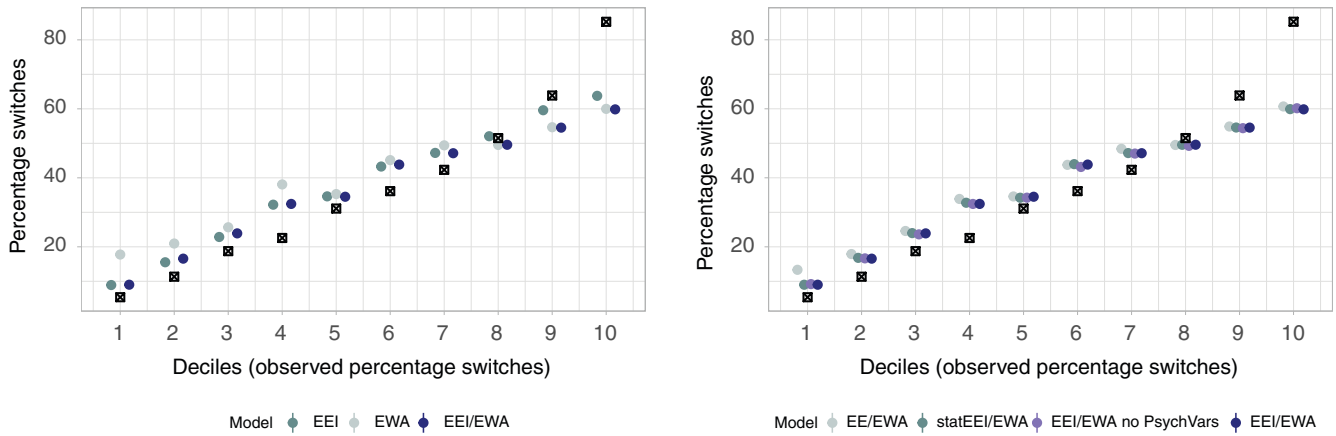
underpredicts the amount of switching for subjects in the last two deciles of the observed percentage of switches, with subjects switching between arms in more than 50% of the rounds. Looking at the performance of the main components of the model, the EWA model does not adjust sufficiently to subjects' behavior, and tends to overpredict the amount of switching for the first few deciles, and underpredicts the percentage of switches for subjects who often switch between arms. This is reflected in the large mean absolute error, at 0.118, compared with 0.092 for the EI/EWA model. The EI model predicts well the percentage of switches between arms, and accommodates particularly well the learning patterns in the last two deciles, characterized by a lot of switching. Its mean absolute error is lowest across models, at 0.078. The key feature benchmark models perform similarly to the proposed model. The EI/EWA model tends to overpredict the amount of switching for subjects in the first few deciles, but performs well for subjects with a moderate to high amount of switching.

6.2.2. Out-of-Sample Predictive Performance. We validate the predictive performance of the proposed model and the various benchmarks using two distinct out-of-sample approaches.

With *temporal forecasting*, we predict a subject's behavior at rounds $[t + 1, T]$, given their history of behavior up to round t (e.g., Ansari et al. 2012, Ascarza and Hardie 2013, Yang et al. 2015, Ascarza et al. 2018). Assessing the ability of our model to predict a subject's choices over time is particularly relevant for our individual-level model, as the model conditions out-of-sample predictions on individual-level parameters and underlying state predictions at time t . We calibrate our proposed model and the relevant benchmarks using the first 80 rounds of data, and use subjects' choices in the last 20 rounds to measure predictive performance.

One limitation of this classic out-of-sample approach is that, in our context, the last 20 rounds are more likely to be geared toward exploitation and inertia than toward exploration. Therefore, the models are not compared on a fully representative sample of exploration/exploitation behaviors.

Figure 12. Posterior Predictive Check: Means and 95% Confidence Bounds of the Observed (Black Crossed Squares) vs. Predicted (Dots) Percentages of Switches Between Arms per Decile



Notes. The left panel shows the proposed model versus main component benchmarks. The right panel shows the proposed model versus key feature benchmarks. The individual-level percentage of switches between arms is a summary measure of the extent of learning a subject engages in. Note that the 95% confidence bounds are very narrow and appear indistinguishable from the means.

To assess the ability of our model to predict the full range of sampling strategies, we also perform *cross-sectional validation*. Implementing a 10-fold cross-validation approach without replacement, we calibrate a model on 90% of the data set (i.e., 80 or 81 subjects, depending on the fold) and predict the full history of behavior for the remaining 10% of subjects in the holdout data. In this exercise, we predict subjects’ behavior from the very first to the very last round.

This latter, *cross-sectional* approach is not without limitations. In particular, it cannot exploit individual-level differences in behavior and tends to predict *average* behaviors, remaining oblivious to some idiosyncratic extreme behaviors observed in the data set.

Interpreting the results from both approaches gives a clear picture of the overall out-of-sample predictive abilities of the EEI/EWA model and its variants.

We use out-of-sample performance measures similar to those used to analyze in-sample fit and posterior predictive performance. Out-of-sample hit rates, MAEs, and MSEs are computed similarly to their

in-sample counterparts using the holdout data for temporal and cross-sectional validation.²¹ Comparing the in-sample fit statistics and the two out-of-sample prediction exercises will inform us on the importance of understanding individual-level exploration/exploitation trade-offs.

Overall, results in Table 5 and Figures 13 and 14 show that although no model uniformly outperforms all others in terms of prediction performance out-of-sample, different specifications are particularly suited to accommodate certain types of behavior. To understand further which specifications are best suited to predict certain behaviors, we investigate how different models fit the behavior of subjects as a function of the extent of switching between arms, similar to the posterior predictive analysis presented in Section 6.2.1.

Comparing the in-sample fit measures to the cross-sectional out-of-sample performance measures, we conclude that ignoring individual heterogeneity hurts prediction performance, with hit rates decreasing by about 10% across all models and measures, whereas the error measures double out-of-sample. Figures 13

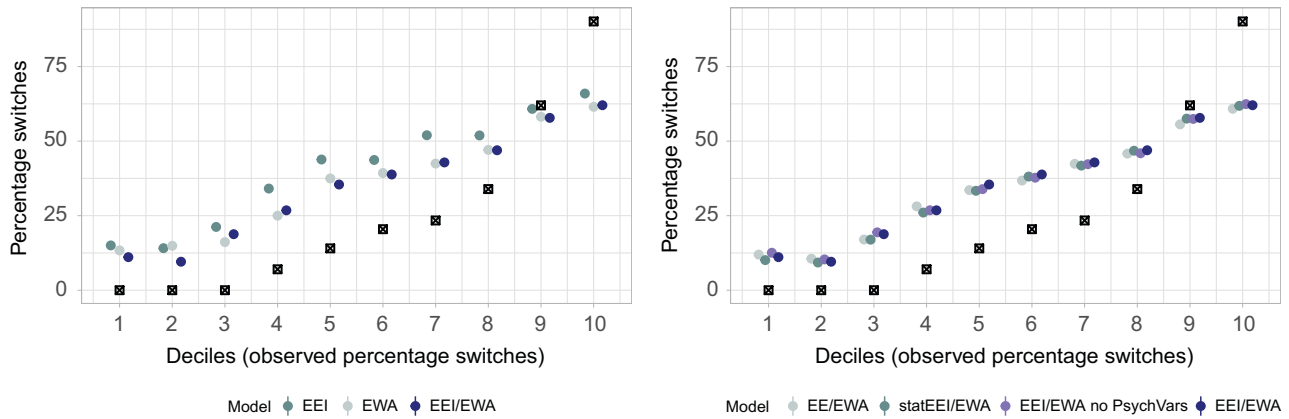
Table 5. Out-of-Sample Prediction Performance

Model	Temporal forecasting ^a			Cross-sectional validation ^b		
	MSE	Hit rates (%)	MAE	MSE	Hit rates (%)	MAE
EEI/EWA (proposed model)	0.176	70.7	0.183	0.293	57.4	0.184
EWA	0.164	69.8	0.192	0.281	53.7	0.226
EEI	0.254	66.0	0.222	0.351	55.4	0.190
stat EEI/EWA	0.174	71.1	0.176	0.300	57.6	0.176
EE/EWA	0.184	70.1	0.185	0.287	59.1	0.191
EEI/EWA no PsychVars	0.180	70.7	0.184	0.296	57.5	0.180

^aThe temporal forecasting performance measures reported here are based on a calibration data set that includes 80 rounds.

^bThe cross-sectional validation measures are based on a 10-fold cross-validation exercise, where subjects in the validation data set are sampled without replacement.

Figure 13. Temporal Out-of-Sample Prediction Performance: Means and 95% Confidence Bounds of the Observed (Black Crossed Squares) vs. Predicted (Dots) Percentages of Switches Between Arms per Decile



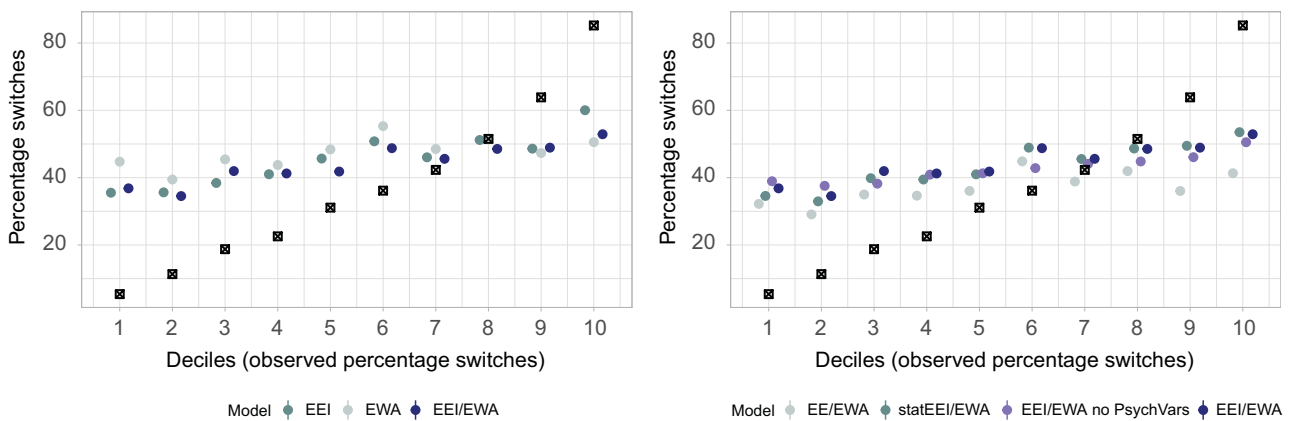
Notes. The left panel shows the proposed model versus main component benchmarks. The right panel shows the proposed model versus key feature benchmarks. The 95% confidence bounds are very narrow and appear indistinguishable from the means.

and 14 reveal a positive correlation between the predicted and the observed percentage of switches between arms across most models, showing that both the proposed model and the various benchmarks are able to accommodate different tendencies in switching behavior. This correlation is lowest in Figure 14, which is not surprising, as we predict subjects' full history of choices out-of-sample, thus ignoring unobserved heterogeneity. "Typical" patterns of behavior where subjects switch between arms in about 30%–50% of the rounds have the lowest error in predicting behavior. Differences between the observed and predicted percentages are highest at the extreme, where subjects switch between arms in either less than 20% of the rounds or in more than 60% of the rounds.

We now discuss the predictive performance of our proposed model against the main components

benchmarks. With reference to Table 5, the proposed model outperforms the EWA model in terms of hit rates and MAEs, but performs worse in terms of MSEs. Looking at Figure 13, the EWA model overpredicts the amount of switching between arms for subjects who consistently choose the same arm in the last 20 rounds of the bandit experiment (first two deciles). Figure 14 shows that the EWA model is not sufficiently flexible to accommodate different types of behavior and predicts a similar amount of switching between arms irrespective of subjects' various patterns of behavior. The EEI model underperforms out-of-sample on average across all measures. It tends to overpredict the amount of switching between arms in the last 20 rounds (see Figure 13). Because of this, the model performs well when predicting the behavior of those who tend to switch frequently between arms even at the later stages

Figure 14. Cross-Sectional Out-of-Sample Prediction Performance: Means and 95% Confidence Bounds of the Observed (Black Crossed Squares) vs. Predicted (Dots) Percentages of Switches Between Arms per Decile



Notes. The left panel shows the proposed model versus main component benchmarks. The right panel shows the proposed model versus key feature benchmarks. The 95% confidence bounds are very narrow and appear indistinguishable from the means.

of the bandit experiment. We observe a similar trend for this model when predicting the full history of choices out-of-sample (see Figure 14).

Last, we discuss the predictive performance of our proposed model against the key feature benchmarks. The EE/EWA model performs worse than the proposed model when predicting the last 20 choices of subjects. It performs better when predicting out-of-sample the subjects' full history of choices, with a higher hit rate and lower MSE compared with the proposed model. However, the MAE of the EE/EWA model is higher than the MAE of the proposed model. With reference to Figure 14, the EE/EWA model predicts well the behavior of those who do not frequently switch between arms, but severely underpredicts the extent of switching for those who do so. The model does not adapt well when predicting out-of-sample the full history of choices. Its predictions only weakly follow the increasing trend in the observed percentage of switches.

The statEEI/EWA slightly outperforms the proposed model on most out-of-sample measures, with the one exception being its higher MSE, at 0.300 versus 0.293 for the proposed model, when predicting choices cross-sectionally. Looking at Figure 14, the statEEI/EWA model performs similarly to the proposed model for "typical" behaviors characterized by moderate switching, and performs slightly better when predicting more extreme behaviors.

The psychometric traits appear to be a good predictor of learning behavior when used for temporal forecasting. The hit rate and MAE of the proposed model are similar to those of the model without psychometrics, but the proposed model has a lower MSE. When predicting the full history of choices out-of-sample, the proposed model performs better in terms of MSE, but worse in terms of MAE, compared with the model without psychometrics (see Table 5). The model without psychometric traits can accommodate "typical" behaviors well for subjects who switch between arms in 15 to 40% of the rounds (see Figure 14). However, the model excluding psychometric traits performs worse than the proposed model when predicting more extreme behaviors. In other words, psychometric variables are not a necessary model component to predict the "central tendencies" of the population, but are valuable to detect and predict extreme behaviors. In the analysis reported in Web Appendix Section WA7, we show that more extreme behaviors correspond to extreme financial (sub)performance; therefore, the model with psychometric variables is particularly relevant from a managerial point of view.

Overall, the proposed model performs well for "typical" learning behaviors and is sufficiently flexible to accommodate extreme behaviors. Psychometrics are instrumental in allowing our model to adapt to more extreme learning tendencies.

Taken together, the in-sample fit measures and the out-of-sample predictions provide converging evidence of the goodness-of-fit of our proposed model to explain learning behavior, and is reasonably robust in predicting various learning patterns. This lends support to our conjecture that a dynamic model accounting for subjects' belief-updating process and allowing for changes over time in subjects' sampling strategies and in their sensitivity to rewards is necessary to understand learning behavior.

7. Discussion

Managers routinely make decisions where they need to strike a healthy balance between exploration and exploitation. Although the ability to search effectively is one of the keys to successful decision making, to the best of our knowledge, research has not yet examined how managers solve such dynamic resource allocation problems. In this paper, we formally examined decision makers' trade-offs between exploration and exploitation strategies, and linked their learning behavior to relevant psychological traits. To elicit learning patterns, we used an experimental design involving a three-armed bandit problem. The experiment mirrored a business setting in which decision makers learned about the profitability of three options, while maximizing overall rewards over time. Decision makers showed an intuitive sense for the basic principles of learning; they sampled the available options multiple times, and eventually repeatedly used the one believed to have the highest expected payoff.

We inferred the optimal sequential sampling strategy in the experimental task using Gittins indices, a canonical result in the stochastic sequential decision-making literature. The behavior of most decision makers is shown to be far from optimal, with a tendency to overexplore options, resulting in rates of switching among options that significantly exceed those warranted by an optimal policy. Such departures from optimality can result in payoffs being forfeited. Decision makers leave money on the table, forgoing on average over 30% of potential revenue.

To capture the dynamics in how decision makers use different sampling strategies and to understand better the underlying mechanisms at play, we used a behavioral model based on a Markovian structure. Decision makers transitioned between exploration and exploitation behaviors, or simply reinforced their previous choice of alternative. Factors such as outcomes of previously sampled alternatives impact decision makers' transitions between sampling strategies. Outcomes that were poorer than expected had a large impact on the choice of sampling strategy, as decision makers tended to shy away from options that gave such disappointing results. This impaired learning

about the distribution of rewards for the disappointing options and contributed to suboptimal sampling.

Although decision makers' trade-offs between exploration and exploitation are suboptimal, we demonstrate that these trade-offs can be explained by looking at their psychometric traits. Individual predictors included risk preferences, decision-making style, and tendencies to maximize or satisfice.

If managers' learning patterns can be anticipated, this would allow the right leader to be appointed for a specific task. Should the situation demand for extended initial exploration of alternative policies, our findings suggest that a more risk-averse decision maker would be best suited to the task. In our experiment, individuals employing a more analytical decision-making style discounted previous rewards and experience of an option more similarly compared with less analytical decision makers. Thus, analytical managers would likely be more appropriate to handle tasks that require constant updating of the expected rewards. Our results also show that, in their search for the best option, maximizers tended to be more sensitive to the expected payoffs of each available option. Managers whose tendency is to maximize are thus likely to be suited to tackle tasks requiring a systematic approach.

We build on the literature on reinforcement learning. Our focus is on specifying a behavioral model that allows us to understand individual-level exploration/exploitation trade-offs. We generalize models that assume underlying learning rules (Camerer and Ho 1999, Nevo and Erev 2012, Erev and Haruvy 2015), and use a hidden Markov model to accommodate several sampling strategies (Ansari et al. 2012). We also compare actual and optimal learning paths. Whereas optimal sequential sampling theory does not offer particularly suitable models to describe decision makers' learning strategies and choice behavior, a normative path offers a useful benchmark to formalize systematic departures from optimal behavior. The computation of the normative path used to benchmark behavior in previous literature is based on the Gittins index specification, which assumes risk neutrality. An interesting avenue for further research would be to study how the predictions of the optimal path change when accounting for decision makers' risk preferences. This can be achieved by using a Whittle index to compute the optimal path, in line with work by Shahrokhi Tehrani and Ching (2019) and Lin et al. (2015). Failing to integrate risk preferences when computing the optimal path leads to an underestimation of the extent of optimal exploration. As a result, the gap between actual rewards and those achieved under optimal sampling might be overestimated. Further studies could strive to understand for which types of decision makers, in terms of their risk profiles, such issues are most pronounced.

We also contribute to the work on managerial decision making (Amaldoss et al. 2000, Ho et al. 2006, Goldfarb and Yang 2009, Goldfarb and Xiao 2011, Goldfarb et al. 2012). We relax assumptions of rational behavior, in an attempt to understand more about managerial learning, and to anticipate decision makers' learning behavior by looking at their psychometric traits.

Our research offers a starting point for future work aimed at providing recommendations to nudge managers toward optimal learning. One area for such recommendations lies in the design of the learning environment. In Web Appendix Section WA1, we investigate how changes in the learning environment can impact decision makers' exploration/exploitation trade-offs and focus on a key feature: the decision time frame. Study 2 allows for repeated opportunities to learn. Study 3 manipulates the planning horizon of the learning environment. We find that offering repeated opportunities to learn and increasing the planning horizon are beneficial, bringing decision makers closer to the optimal path. Although the additional studies are a reasonable starting point, many other features of the learning environment should be investigated, and our work opens fruitful avenues for further research. As an example, researchers could study the interaction between the learning patterns of individual managers and the business culture in which they operate. How would an overexplorer behave in an environment in which the focus was on either experimentation and innovation or on efficiency and strategy refinement? Using an experimental design in which managers are encouraged to engage in either an exploration or an exploitation strategy could provide insights into how their performance is affected by a match or mismatch between inherent and induced learning styles.

Further studies could attempt to devise toolboxes, designed to compensate individual tendencies to either over- or underexplore or to limit the extensive discounting of previously acquired information.²² We have provided the basics for prescriptive analysis, by modeling learning at the individual level, and showing how researchers can systematically investigate changes in behavior across different experimental conditions. Our analysis links learning tendencies to psychological traits, which allows us to anticipate managers' trade-offs between exploration and exploitation.

We encourage future research to look more deeply at managers' learning tendencies and at how they impact organizational behavior and profitability, a promising area that has not yet been sufficiently examined.

Acknowledgments

The authors thank the senior editor, associate editor, and two anonymous reviewers for their constructive comments, which improved this paper. They gratefully acknowledge the helpful comments from Gui Liberali; Ayse Oncüler;

participants at seminars at Erasmus University, Rotterdam School of Management, Humboldt-Universität zu Berlin, McGill University, the National University of Singapore, the University of Chicago, and the University of Maryland; and participants at the Behavioral Industrial Organization and Marketing Conference, the Carnegie School of Organizational Learning, and the INFORMS Marketing Science Conference. They thank the Erasmus Research Institute of Management for providing financial support, and the Dutch national e-infrastructure with support of the SURF Cooperative. This paper is based on the first author's doctoral dissertation at ESSEC Business School.

Endnotes

¹ Further information is available on Google's support website (see <https://cloud.google.com/blog/products/ai-machine-learning/how-to-build-better-contextual-bandits-machine-learning-models>; accessed on March 19, 2021).

² See https://help.optimizely.com/Build_Campaigns_and_Experiments/Stats_Accelerator (accessed on March 19, 2021).

³ We thank an anonymous reviewer for suggesting the additional studies.

⁴ Bandit experiments with a smaller number of rounds and similar average rewards per arms are likely to lead to random switching, and to reduce the discriminatory power of any descriptive model (Rapoport and Budescu 1997, Gans et al. 2007).

⁵ We use the classic power utility function $u(x) = x^r$ and elicit the CRRA parameter r . Values of r below one imply risk-averse behavior, and values above one imply risk-seeking behavior. See details in Web Appendix Section WA2.

⁶ Given the specificity of the task, we expected to find a higher average score on the analytical subscale than on the intuitive subscale.

⁷ The average payment was calibrated to the hourly wage standards imposed by the experimental laboratory.

⁸ Further details on the computation can be found in Web Appendix Section WA3.

⁹ Subjects' intertemporal preferences can differ significantly, and optimal learning is dependent on the level of discounting of future rewards. Although we cannot rule out definitively the impact of time preferences on subjects' choices, several features of the problem mitigate this issue. Because we are studying managerial learning, the weight on long-term versus short-term rewards is part of a firm's strategy, and is thus less impacted by individual preferences. Our problem is a finite horizon problem, and subjects are paid in proportion to the overall rewards accumulated during the task. It is unlikely that subjects strongly discounted future rewards. Moreover, in our behavioral model and in the robustness analysis in Web Appendix Section WA4, we account for the extent to which subjects forget earlier outcomes and rely on recent outcomes when making choices. This reveals interesting learning patterns, which we discuss in Section 6 and in Web Appendix Section WA4.

¹⁰ Our study involves choices over a finite horizon. In order to implement an "infinite" horizon in our experiment, we could use a probabilistic end rule (Gans et al. 2007). The experimental task can end at every round, with a fixed and known probability. The probability is used to discount expected rewards. Banks et al. (1997) note some concerns with this procedure. First, subjects may not have a good understanding of this probabilistic rule. Second, the small chance that the experiment could last for several hours is not believable. Also, in our context, managers cannot afford to test prospective actions over an infinite time horizon. We therefore used a 100-day finite horizon. Studies 2 and 3 in Web Appendix Section

WA1, speak to how exploration/exploitation trade-offs change over the planning horizon and with repeated opportunities to learn.

¹¹ In Web Appendix Section WA4, we present a robustness analysis to this specification, where previous rewards are decayed before entering the expected value computation, with several decay levels. A specification that heavily decays previous rewards before entering the expected value computation performs best in-sample, whereas the out-of-sample prediction measures give a slight edge to the specification presented here. We thank an anonymous reviewer for suggesting this analysis.

¹² The problem of how subjects form prior beliefs is fundamentally different than understanding how decision makers learn. Here we do not have sufficient experimental variation to identify subjects' prior beliefs, nor the degrees of freedom to estimate them. This is why we rely on the information we presented to subjects in the experimental instructions to set such beliefs and let future research investigate the process of forming prior beliefs.

¹³ The decay parameters are bounded between zero and one, and we apply an inverse-logit transformation. Both sensitivity parameters $\lambda^{Explore}$ and $\lambda_i^{Exploit}$ are bounded between zero and one. We apply an inverse logit and an exponential transformation to ensure the proper ordering of the sensitivity parameters. Given the scale of the attractions of options, a value of one corresponds to a quasi-deterministic choice rule. Our estimates of the sensitivity parameters are significantly below one across all models.

¹⁴ We use the following hyperpriors: $vec(\gamma) \sim \mathcal{N}(0, 1)$, $\Omega \sim \mathcal{LKJ}(2)$, $vec(\tau) \sim Exponential(1)$. At the aggregate level, $\beta_{1kk'} \sim \mathcal{N}(0, 1)$.

¹⁵ We conducted a simulation study to test whether our model parameters are empirically identified, and found that parameters are well recovered for simulated data similar to our data set from Study 1. See Web Appendix Section WA6 for details.

¹⁶ The Gelman and Rubin (1992) statistic is below the acceptable threshold of 1.1 for all model parameters, showing that the chains have converged to the stationary distributions.

¹⁷ Figures WA6, WA7, and WA8 in Web Appendix Section WA7 show significant heterogeneity in the baseline propensities of subjects to switch between states and in the parameters governing their belief-updating process.

¹⁸ We report the means and 95% HDIs of these parameters in Web Appendix Section WA7.

¹⁹ The higher the CRRA parameter, the more risk seeking a decision maker is. Interestingly, risk-seeking decision makers seem more likely to return to exploration from either exploitation or inertia, although the parameters estimates did not reach statistical significance ($\delta_{CRRA, \beta_{0i, 21}} = 0.945$, 95% HDI = $[-0.264, 2.128]$; $\delta_{CRRA, \beta_{0i, 31}} = 0.674$, 95% HDI = $[-0.462, 1.856]$).

²⁰ In Web Appendix Section WA8, we present a post hoc exploratory analysis highlighting the link between various psychological profiles, learning tendencies, and overall rewards.

²¹ Psychometrics are integrated in the calibration data set used for the temporal out-of-sample predictions in line with the model specified in Section 5.6, as we estimate individual-level parameters and use these parameters for temporal forecasting. When predicting the full history of choices out-of-sample, we use the full data set to first compute the means of the psychometric traits, and then mean-center the psychometrics of the subjects in the calibration and validation samples relative to these means. Therefore, we keep the same mean psychometric traits across all 10 folds used for cross-validation as a proxy for the stable population-level traits. This reduces extraneous variation in parameter estimates across the 10 folds and facilitates model comparison.

²² To facilitate further studies, the data set and codes used in this research are available at https://github.com/alinafere/manAGERIAL_explORATION_explOITATION_tradeOFFS.

References

- Adler PS, Goldoftas B, Levine DI (1999) Flexibility vs. efficiency? A case study of model changeovers in the Toyota production system. *Organ. Sci.* 10(1):43–68.
- Ahn WY, Vasilev G, Lee SH, Busemeyer JR, Kruschke JK, Bechara A, Vassileva J (2014) Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. *Front. Psychol.* 5:849.
- Amaldoss W, Meyer RJ, Raju JS, Rapoport A (2000) Collaborating to compete. *Marketing Sci.* 19(2):105–126.
- Anderson CM (2001) Behavioral models of strategies in multi-armed bandit problems. Unpublished PhD Thesis, California Institute of Technology, Pasadena.
- Ansari A, Montoya R, Netzer O (2012) Dynamic learning in behavioral games: A hidden Markov mixture of experts approach. *Quant. Marketing Econom.* 10(4):475–503.
- Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Sci.* 32(4):570–590.
- Ascarza E, Netzer O, Hardie BGS (2018) Some customers would rather leave without saying goodbye. *Marketing Sci.* 37(1):54–77.
- Baardman L, Fata E, Pani A, Perakis G (2019) Learning optimal online advertising portfolios with periodic budgets. Preprint, submitted March 27, <http://dx.doi.org/10.2139/ssrn.3346642>.
- Banks J, Olson M, Porter D (1997) An experimental analysis of the bandit problem. *Econom. Theory* 10(1):55–77.
- Benner MJ, Tushman M (2002) Process management and technological innovation: A longitudinal study of the photography and paint industries. *Admin. Sci. Quart.* 47(4):676–706.
- Benner MJ, Tushman ML (2003) Exploitation, exploration, and process management: The productivity dilemma revisited. *Acad. Management Rev.* 28(2):238–256.
- Betancourt MJ, Girolami M (2015) Hamiltonian Monte Carlo for hierarchical models. Upadhyay SK, Singh U, Dey DK, Loganathan A, eds. *Current Trends in Bayesian Methodology with Applications* (CRC Press, Boca Raton, FL), 79–100.
- Biele G, Erev I, Ert E (2009) Learning, risk attitude and hot stoves in restless bandit problems. *J. Math. Psych.* 53(3):155–167.
- Busemeyer JR, Stout JC (2002) A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psych. Assessment* 14(3):253–262.
- Camerer C, Ho TH (1999) Experience-weighted attraction learning in normal form games. *Econometrica* 67(4):827–874.
- Camerer CF, Ho TH, Chong JK (2004) A cognitive hierarchy model of games. *Quart. J. Econom.* 119(3):861–898.
- Cohen JD, McClure SM, Yu AJ (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. Roy. Soc. London Ser. B* 362(1481):933–942.
- Crosetto P, Filippin A (2013) The “bomb” risk elicitation task. *J. Risk Uncertainty* 47(1):31–65.
- Cui TH, Mallucci P (2016) Fairness ideals in distribution channels. *J. Marketing Res.* 53(6):969–987.
- Daw ND, O’Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441(7095):876–879.
- Denrell J (2005) Why most people disapprove of me: Experience sampling in impression formation. *Psych. Rev.* 112(4):951–978.
- Denrell J (2007) Adaptive learning and risk taking. *Psych. Rev.* 114(1):177–187.
- Denrell J, March JG (2001) Adaptation as information restriction: The hot stove effect. *Organ. Sci.* 12(5):523–538.
- Erev I, Haruvy E (2005) Generality, repetition, and the role of descriptive learning models. *J. Math. Psych.* 49(5):357–371.
- Erev I, Haruvy E (2015) Learning and the economics of small decisions. *The Handbook of Experimental Economics*, vol. 2 (Princeton University Press, Princeton, NJ), 638–716.
- Erev I, Roth AE (2014) Maximization, learning, and economic behavior. *Proc. Natl. Acad. Sci. USA* 111(3):10818–10825.
- Erev I, Ert E, Yechiam E (2008) Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *J. Behav. Decision Making* 21(5):575–597.
- Gans N, Knox G, Croson R (2007) Simple models of discrete choice and their performance in bandit experiments. *Manufacturing Serv. Oper. Management* 9(4):383–408.
- Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.* 7(4):457–472.
- Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Statist. Comput.* 24(6):997–1016.
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian Data Analysis* (CRC Press, Boca Raton, FL).
- Gilboa I, Pazgal A (2001) Cumulative discrete choice. *Marketing Lett.* 12(2):119–130.
- Gittins J, Glazebrook K, Weber R (2011) *Multi-Armed Bandit Allocation Indices*, 2nd ed. (Wiley, Hoboken, NJ).
- Gittins JC, Jones DM (1979) A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika* 66(3):561–565.
- Goldfarb A, Xiao M (2011) Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets. *Amer. Econom. Rev.* 101(7):3130–3161.
- Goldfarb A, Yang B (2009) Are all managers created equal? *J. Marketing Res.* 46(5):612–622.
- Goldfarb A, Ho TH, Amaldoss W, Brown AL, Chen Y, Cui TH, Galasso A, et al. (2012) Behavioral models of managerial decision-making. *Marketing Lett.* 23(2):405–421.
- Hastie T, Tibshirani R, Friedman J (2016) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York).
- Hauser JR, Liberali GG, Urban GL (2014) Website morphing 2.0: Switching costs, partial exposure, random exit, and when to morph. *Management Sci.* 60(6):1594–1616.
- Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. *Marketing Sci.* 28(2):202–223.
- Hill DN, Nassif H, Liu Y, Iyer A, Vishwanathan SVN (2017) An efficient bandit algorithm for realtime multivariate optimization. Matwin S, Yu S, Farooq F, eds. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 1813–1821.
- Ho TH, Lim N, Camerer CF (2006) Modeling the psychology of consumer and firm behavior with behavioral economics. *J. Marketing Res.* 43(3):307–331.
- Horowitz A (1975) Experimental study of the two-armed bandit problem. Unpublished PhD thesis, University of North Carolina, Chapel Hill.
- Lattimore T (2016) Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. Feldman V, Rakhlin A, Shamir O, eds. *Proc. Machine Learn. Res. Conf. Learn. Theory*, vol. 49 (PMLR, Columbia University, New York), 1–32.
- Liberali G, Ferecatu A (2019) Morphing consumer dynamics: Bandits meet HMM. Preprint, submitted December 16, <http://dx.doi.org/10.2139/ssrn.3495518>.
- Lin S, Zhang J, Hauser JR (2015) Learning from experience, simply. *Marketing Sci.* 34(1):1–19.
- March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.
- March JG (1996) Learning to be risk averse. *Psych. Rev.* 103(2):309–319.
- Meyer RJ, Shi Y (1995) Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Sci.* 41(5):817–834.

- Misra K, Schwartz EM, Abernethy J (2019) Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Sci.* 38(2):226–252.
- Murphy KP, Bach F (2012) *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA).
- Nenkov GY, Morrin M, Ward A, Schwartz B, Hulland J (2008) A short form of the Maximization Scale: Factor structure, reliability and validity studies. *Judgment Decision Making* 3(5):371–388.
- Nevo I, Erev I (2012) On surprise, change, and the effect of recent outcomes. *Front. Psych.: Cognitive Sci.* 3:24.
- Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* 32(2):551–562.
- Novak TP, Hoffman DL (2009) The fit of thinking style and situation: New measures of situation-specific experiential and rational cognition. *J. Consumer Res.* 36(1):56–72.
- Posen HE, Levinthal DA (2011) Chasing a moving target: Exploitation and exploration in dynamic environments. *Management Sci.* 58(3):587–601.
- Rapoport A, Amaldoss W (2000) Mixed strategies and iterative elimination of strongly dominated strategies: an experimental investigation of states of knowledge. *J. Econom. Behav. Organ.* 42(4):483–521.
- Rapoport A, Budescu DV (1997) Randomization in individual choice behavior. *Psych. Rev.* 104(3):603–617.
- Rescorla R, Wagner A (1972) A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. Abraham HB, William FP, eds. *Classical Conditioning: Current Research and Theory* (Appleton-Century-Crofts, New York), 64–99.
- Roth AE, Erev I (1995) Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econom. Behav.* 8(1):164–212.
- Roth AE, Erev I (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Amer. Econom. Rev.* 88(4):848–881.
- Sarin R, Vahid F (1999) Payoff Assessments without probabilities: A simple dynamic model of choice. *Games Econom. Behav.* 28(2):294–309.
- Schwartz B, Ward A, Monterosso J, Lyubomirsky S, White K, Lehman DR (2002) Maximizing vs. satisficing: Happiness is a matter of choice. *J. Personality Soc. Psych.* 83(5):1178–1197.
- Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Sci.* 36(4):500–522.
- Scott SL (2010) A modern Bayesian look at the multi-armed bandit. *Appl. Stochastic Models Bus. Indust.* 26(6):639–658.
- Shahrokhi Tehrani S, Ching AT (2019) A heuristic approach to explore: The value of perfect information. Preprint, submitted May 21, <http://dx.doi.org/10.2139/ssrn.3386737>.
- Shmueli G (2010) To explain or to predict? *Statist. Sci.* 25(3):289–310.
- Simon HA (1959) Theories of decision-making in economics and behavioral science. *Amer. Econom. Rev.* 49(3):253–283.
- Steyvers M, Lee MD, Wagenmakers EJ (2009) A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psych.* 53(3):168–179.
- Toplak ME, West RF, Stanovich KE (2011) The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory Cognition.* 39(7):1275–1289.
- Tushman ML, O'Reilly CA (1996) Ambidextrous organizations: Managing evolutionary and revolutionary change. *California Management Rev.* 38(4):8–29.
- Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4):297–323.
- Urban GL, Liberali GG, MacDonald E, Bordley R, Hauser JR (2014) Morphing banner advertising. *Marketing Sci.* 33(1):27–46.
- Whittle P (1988) Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.* 25:287–298.
- Yang LC, Toubia O, De Jong MG (2015) A bounded rationality model of information search and choice in preference measurement. *J. Marketing Res.* 52(2):166–183.