Beyond the Exploration–Exploitation Trade-off: The Behavioral Structure of Switching between Alternatives*

Jordi Martinez-Muñoz†

November 1, 2026
Click here for the most recent version

Abstract

Many sequential decisions, such as research projects, investments and partnerships, involve choosing among alternatives whose quality is learned only through experience. Standard models predict that people switch between alternatives only when they learn something new. I test this in an incentivized one-armed bandit experiment: subjects choose repeatedly between a familiar option and an uncertain one, a reward reveals that the uncertain option is better, and outcomes of the familiar option reveal nothing. The model predicts single switching: try the uncertain option until it pays off or enough failures pile up, then stay with one option for good. About half of subjects return to the uncertain option after abandoning it with no new information about it, and 39% leave it at least once even after learning it is the better one. Those who switch this way explore inferior options longer, find good ones later, and earn 20 to 25% less than single-switchers. Part of this switching is a reaction to outcomes that carry no information, such as rewards from the familiar option. When subjects can choose plans that limit their reaction to outcomes, about half of those who deviated adopt single switching, consistent with noisy execution of a correct plan. The rest choose plans that keep switching, consistent with an incorrect model of the problem or non-standard preferences.

Keywords: Bandit problems, Exploration–exploitation trade-off, Sequential experimentation, Switching, Choice architecture, Experiment
JEL classification: C91, D81, D83, D91.

*I am indebted to my advisor Emanuel Vespa, and to Joel Sobel and Isabel Trevino, for their guidance and support on this project. I also thank Botond Kőszegi, Philippos Louis, Marta Serra-Garcia and Denis Shishkin for valuable conversations and comments, as well as participants in the Theory-Behavioral-Experimental Graduate Seminar at UC San Diego. I am grateful to all study participants for their time and patience. This research was funded in part by a UC San Diego Department of Economics research grant and a grant from the Yankelovich Center for Social Science Research at UC San Diego. The study was approved by the UC San Diego IRB under protocol #810859 and pre-registered on AsPredicted as #246452 (Session 1), #301151 (Session 2) and #306498 (Session 3). All errors are my own.

†Job Market Paper. University of California San Diego, Department of Economics. Email: jom001@ucsd.edu.

1 Introduction

Many sequential decisions involve alternatives whose quality is revealed only through experience. Consider a researcher who, every month, works on either a familiar project, whose payoff distribution she knows, or an unknown project, which may be better or worse. She learns about the unknown project only by working on it, at the cost of the payoff the familiar project could have given her. Bandit problems formalize this exploration–exploitation trade-off (Robbins, 1952),[1] and their solution ties switching to information: an agent should change alternatives only when what she has learned changes their relative value.

This paper tests that link experimentally, in the one-armed bandit with conclusive good news, that is, one in which a reward fully reveals the option's quality. In the researcher example, if she chooses the unknown project one month and it succeeds, she will know that the project is good and better than the familiar one. The solution has a simple structure, single switching: the researcher works on the unknown project until it succeeds or until enough failures accumulate, then stays with one project for good. Two informational restrictions deliver it. Working on the familiar project reveals nothing about the unknown one, so once she prefers the familiar project, the unknown project never becomes more attractive and she never returns to it. And a good unknown project first-order stochastically dominates the familiar one, so once the unknown project succeeds, she never leaves it. Any other switch has no informational justification. I find that people's switching between options does not track information: it responds to outcomes that carry no information, even once the options' quality is known.

Bandit experiments have focused almost exclusively on behavior while the quality of at least one option (arm in the bandit literature) is unknown, what I call the Quality Uncertain (QU) stage of the problem (see Hudja and Woods, 2024b, for a review). However, because I take a step back and instead study the informational content of decisions, I also need to study behavior in the Quality Certain (QC) stage of the problem, after the unknown arm pays. This extension allows me to ask whether potential deviations from single switching are exclusive to uncertainty about arm quality or not.

My experiment implements the problem with jars and balls. In each of the 30 rounds of a problem, subjects draw a ball from one of two jars. The familiar jar pays 50 points with probability 0.1. The unknown jar is good or bad with equal probability, where a good jar pays with probability 0.2 and bad jars never pay. In the baseline treatment, subjects

[1]For the infinite-horizon discounted problem, Gittins and Jones (1974) characterized the optimal policy using what is now known as the Gittins index.

choose jars freely; later treatments vary only when and how subjects can change jars. The pre-registered experiment involved 300 subjects.

First, I find that neither changes in information nor uncertainty about arm quality is necessary for switching, as over-switching (deviations from the single-switching rule) is widespread in both stages. About half of subjects return to the unknown arm after switching away from it in the QU stage. In the QC stage, 39% of subjects switch away from it at least once. Over-switching thus falls once quality is known, but it persists, and over-switching is correlated across the two stages ($\rho = 0.41$, $p < 0.001$).

Second, in particular, over-switchers respond in both stages to outcomes that carry no information. In the QU stage, a second reward within the current wave (a sequence of consecutive choices of the same arm) causally reduces the probability of switching away by about 10 percentage points, and once quality is known (QC), even a single reward from the unknown arm reduces it by 6.0 percentage points. This response alone rules out random switching and other models in which choices ignore outcomes. Further analysis allows me to rule out win-stay/lose-shift, reinforcement learning and treating the familiar arm as of unknown quality, and to rule out utility curvature with wealth effects as the leading reason for over-switching.

Third, heterogeneity among over-switchers reveals two potential mechanisms that operate in both stages. Over-switching may reflect a noisy execution of a correct plan, triggered by the experience of playing: random errors, emotional reactions, or boredom; or it may instead reflect a valuation distortion: incorrect models of the problem or non-standard preferences.

Two treatments limit how far switching can respond to realized outcomes, which should correct noisy execution but not a valuation distortion. Under Block Choice, subjects choose an arm together with the number of rounds to draw from it, so they decide when to reconsider before seeing the outcomes. Under Ex Ante Rules, they choose an automatic rule before the problem starts and see only the final score. In the QU stage, Block Choice moves about half of the over-switchers to single switching, and about as many choose a single-switching rule under Ex Ante Rules; about a quarter keep over-switching under both. The treatments have similar effects in the QC stage, where subjects tend to respond the same way to both.

Importantly, subjects consistent with each mechanism also differ in their free-switching behavior in the QU stage: those closer to noisy execution switch less often and are less likely to switch right after a failure, and they are better predicted by a single-switching plan with occasional slips than by models without that plan, while those closer to a valuation

distortion are not.

Finally, over-switching has implications for the exploration margin and performance. Over-switchers first abandon the unknown arm earlier than single-switchers, yet explore bad unknown arms about 40% more. When the unknown arm is good, they reach its first reward later, because they interleave draws from the familiar arm. Single-switchers, by contrast, cluster around the risk-neutral Bayesian benchmark of eight rounds of exploration. Over-switchers earn 20 to 25% less. Further treatments that restrict subjects to a single switch, decided round by round or ex ante, cut exploration among the over-switchers who explore most, bringing them closer on average to single-switchers.

Hence, in sequential experimentation, people change options without a corresponding change in information, but this is not exclusive to the need to explore: it persists even when the alternatives are fully known. More generally, people respond to outcomes that carry no information, with heterogeneity in the degree to which this can be fixed by restricting such responses.

These results are important because sequential experimentation is a common problem that agents face. Beyond the researcher's choice of projects, such decision problems arise, for example, in evaluating investment projects, complex portfolio options, political strategies, and partnerships. Field evidence documents costly returns to previously discarded alternatives,[2] but these returns are hard to interpret, since a return may reflect new information about the alternative. Studying the general phenomenon therefore requires the controlled and simplified environment of an experiment.

Connection to the literature. This paper contributes to the literature on the exploration–exploitation trade-off and its canonical representation as the multi-armed bandit problem (Robbins, 1952), with applications throughout economics, including strategic experimentation (Bergemann and Välimäki, 2008).[3] The experimental literature on bandits has examined

[2]In romantic relationships, "churning" couples who break up and reconcile with the same partner report worse mental health and relationship quality outcomes (Halpern-Meekin, Manning, Giordano, and Longmore, 2013; Halpern-Meekin and Turney, 2016, 2023; Monk, Ogolsky, and Oswald, 2018). In labor markets, "boomerang" employees rehired by a former employer perform, on average, no better than external or internal hires and improve less over time (Swider, Liu, Harris, and Gardner, 2017; Arnold, Van Iddekinge, Campion, Bauer, and Campion, 2021). An analogous pattern holds among NBA players who return to a former team (Grohsjean, Dokko, and Yang, 2025). In consumer markets, firms that "win back" defected customers find that reacquired customers disproportionately churn a second time for the very reason that drove their original defection, unless that reason is explicitly resolved (Kumar, Bhagwat, and Zhang, 2015; Kumar, Leszkiewicz, and Herbst, 2018).

[3]Examples of applications of bandits include, but are not limited to, public goods (Keller, Rady, and Cripps, 2005; Keller and Rady, 2010; Hörner, Klein, and Rady, 2022), innovation contests (Bimpikis et al., 2019; Halac et al., 2017), long-term contracts (Halac et al., 2016), moral hazard in teams (Bonatti

both how much subjects experiment and how they respond to incentives and information.

The closest antecedent to this paper is Meyer and Shi (1995), who document that subjects in a one-armed bandit with ambiguity switch far more often than the optimal single-switch policy allows and find hypersensitivity to short-term feedback, which they attribute to a "never stick with a loser" heuristic. They do not report, however, how frequent this behavior is.[4] My data confirm that switching responds to outcomes, causally and even when outcomes carry no information, but not as that heuristic predicts: over-switchers leave the familiar arm right after a reward as often as right after a failure. In other one-armed bandit experiments, the familiar arm pays a deterministic amount. Banks, Olson, and Porter (1997) and Hoelzemann and Klein (2021) document adherence to the predicted experimentation strategy.[5] Hudja and Woods (2024a) impose single switching as part of the experimental design.[6] Other experimental studies consider two-armed environments in which the qualities of both alternatives are never fully revealed, so outcomes after switching are always informative (Anderson, 2001; Gans, Knox, and Croson, 2007; Hudja and Woods, 2025). In all these environments, every deviation happens under quality uncertainty: it can be attributed to how subjects process information about how likely an arm is to be good, or even to its mere presence.[7]

My design lets me test which deviations persist in the QC stage, to what extent, and whether they seem to stem from the same mechanisms. It also connects the bandit problem to two literatures on choice under outcome uncertainty with known probabilities: biased beliefs about random sequences, such as the gambler's fallacy and the hot-hand bias (Benjamin, 2019), and probability matching (Vulkan, 2000). In the QC stage, subjects repeatedly choose between two options with known reward probabilities, as in matching tasks that state the probabilities (West and Stanovich, 2003; Newell and Rakow, 2007; Koehler and James, 2010; Agranov, Healy, and Nielsen, 2023). The design also relates to the description–experience gap (Barron and Erev, 2003; Hertwig, Barron, Weber, and Erev, 2004; Hertwig and Erev, 2009; Wulff, Mergenthaler-Canseco, and Hertwig, 2018). The description of the environment

and Hörner, 2011), and political economy (Strulovici, 2010; Khromenkova, 2015; Bowen, Chan, Dube, and Lambert, 2016; Anesi and Bowen, 2021; Bowen, Hwang, and Krasa, 2022).

[4]Their evidence also rests on average switching across task conditions among about 18 subjects, so it does not establish how widespread over-switching is or whether it characterizes particular subjects.

[5]In Hoelzemann and Klein (2021), this refers to their control treatment, in which the other players' information has no value; frequent switching appears in their strategic treatment.

[6]Hoelzemann and Klein (2021) and Hudja and Woods (2024a) use continuous-time bandits with a random ending.

[7]Meyer and Shi themselves note that subjects may have treated both options as uncertain, in which case switching back and forth is rational.

is complete and specifies the only channel through which experience should matter, the posterior about the unknown arm. Yet experience affects switching through other channels: over-switching falls when subjects cannot respond to outcomes as they arrive (Block Choice) or do not experience them at all (Ex Ante Rules).

I make two further contributions. First, I address a potential problem common to one-armed bandit experiments. Single switching is optimal under risk neutrality, or under non-linear utility over each round's payoff. But in an experiment, payoffs arrive at the end of the session and each choice takes seconds, so subjects may evaluate total earnings instead, and then curvature can break single switching through wealth effects unless absolute risk aversion is constant (CARA). Over-switching could then simply reflect deviations from CARA. I test this explanation directly and rule it out as the leading reason for over-switching. Second, I bring together findings that earlier experiments report separately. Horowitz (1973), the first documented experiment on bandits, finds that subjects in a two-armed bandit oversample less promising options when base rates of success are low (as reported by Meyer and Shi, 1995), and Meyer and Shi (1995) find that subjects switch more at lower base rates and that they under-experiment with promising options and over-experiment with unpromising ones. My data connect these findings through the relation between switching and exploration within subjects.

The rest of the paper is organized as follows. Section 2 presents the model and hypotheses, and Section 3 the experimental design. Section 4 presents the results: switching behavior and its consequences, the mechanisms behind over-switching, and how imposing single switching affects exploration. Section 5 discusses the implications, and Section 6 concludes.

2 Model and Hypotheses

I study a finite-horizon one-armed bandit problem in which a decision maker chooses between an unknown arm (Arm U) and a familiar arm (Arm F) over N periods. In each period, the chosen arm yields either a reward of size $w > 0$ or zero, with rewards accumulating additively across rounds.

Arm F delivers a reward with known probability $\theta_F \in (0, 1)$ in every period. This is the familiar arm since the payoff distribution is known to the DM. Arm U is the unknown arm since it has uncertain quality: with prior probability $p_U \in (0, 1)$, it is *good* and yields a reward with probability $\theta_U \in (0, 1)$; with probability $1 - p_U$, it is *bad* and never yields a reward. The type of Arm U is realized at the beginning of the problem and remains fixed

throughout the horizon.

The arms offer the same reward (w) when successful and differ in the information available about their payoff probabilities: Arm U allows learning about its quality, while Arm F has a known payoff probability.

For the utility function, I assume weak monotonicity and allow for nonlinearities. Because payoffs are realized only at the end of the problem (to match the experimental setting), I assume no discounting and allow the DM to maximize over cumulative payoffs, instead of period-by-period payoffs. That is, I consider $u : \mathbb{R} \to \mathbb{R}$ with $u(Y^{'}) \geq u(Y)$ for $Y' \geq Y$, applied to $Y = \sum_t y_t$, where y_t is the period-t payoff. The decision maker maximizes expected utility, assuming constant absolute risk aversion (CARA). I discuss in Section 4.2 how wealth effects can break the predictions below and the empirical evidence on them.

This environment captures the standard exploration–exploitation tradeoff: sampling Arm U generates potentially profitable information about its quality, while Arm F does not.

Stages of the Decision Problem

For analytical and empirical purposes, I divide each problem into two informational stages.

The *Quality Uncertain (QU) stage* begins in round 1 and continues until the first reward from Arm U is observed, if such an event occurs, or until the end of the problem. During this stage, the decision maker remains uncertain about the quality of Arm U.

The *Quality Certain (QC) stage* activates if Arm U returns a reward. It begins immediately after the first reward from Arm U and continues until the final round. From this point onward, Arm U is known to be good and first-order stochastically dominates Arm F.

If the decision maker commits to Arm F without observing a reward from Arm U, subsequent rounds are formally exploitative but will be classified as part of the QU stage, since learning about Arm U remains possible. This decomposition classifies behavior according to the agent's informational state rather than observed choices and aligns directly with the empirical analysis.

Within this decomposition, I define *exploration* (of Arm U) as the total number of periods in a problem in which Arm U is chosen during the QU stage, counted cumulatively whether or not these periods are consecutive. Choices of Arm F, and choices of Arm U in the QC stage, do not count as exploration.

A *wave* is a sequence of consecutive choices of the same arm; a wave ends when the decision maker switches to the other arm.

Empirical Parameters

Specifically, for the experimental implementation I work with decision problems of 30 rounds each. The symmetric payoff for each arm is 50 points per reward ($w = 50$). The probability of the risky arm being good $p_U = 0.5$, and the probabilities of drawing a reward from each arm are $\theta_U = 0.2$ and $\theta_F = 0.1$ for Arm U and Arm F, respectively.

The 50/50 prior eliminates asymmetric base-rate information, minimizing the scope for base-rate neglect to mechanically influence exploration behavior (see Benjamin, 2019, for a survey).[8] Conditional on yielding a reward, both arms return the same amount, so decisions hinge solely on the relative likelihood of obtaining rewards rather than on payoff differences.

I choose θ values in increments of 0.1 to align with evidence that subjects often discretize beliefs in tens of percentage points in belief-updating tasks (Bland and Rosokha, 2024). This parameterization simplifies inference while making salient that a good Arm U is twice as likely to deliver a reward as Arm F. Since $p_U\theta_U = \theta_F = 0.1$, both arms give the same expected payoff in the first period, but Arm U has a greater value of information because a reward in the first round reveals that the arm is good, in which case it FOSD Arm F. Therefore, under constant absolute risk aversion, playing Arm U in the first period is optimal whatever the degree of risk aversion.

Finally, the finite horizon removes uncertainty about random termination, avoiding the introduction of an additional stochastic element—often used in experiments to approximate infinite-horizon problems (Roth and Murnighan 1978)—beyond uncertainty about arm quality and realized rewards. The chosen horizon balances the need for sufficient scope for exploration and behavioral deviations against practical constraints. Because rewards may arrive early, censoring exploration, multiple problems per subject are required to recover sufficient variation in behavior, which becomes difficult when problems are too long.

Hypotheses

H1 Optimal Switching Path. Under expected-utility maximization with constant absolute risk aversion, the optimal switching path is characterized by three features:

(i) The decision maker starts with Arm U and continues exploring until either a reward is observed or accumulated unfavorable evidence makes switching to Arm F optimal.

[8] A symmetric-prior design to limit base-rate confounds is also used by Fan, Liang, and Peng (2026) in their study of the inference–forecast gap in belief updating.

(ii) If the decision maker switches to Arm F before observing a reward from Arm U, the decision maker never returns to Arm U.

(iii) If Arm U returns a reward, the decision maker stays with Arm U for the remainder of the problem.

Hence, optimal behavior imposes two restrictions on switching behavior. In the QU informational stage, once Arm F is chosen, no information about the underlying quality of Arm U can be received and therefore no new information can provide a reason to switch back to it. In the QC informational stage, no signal can provide a reason to switch away from Arm U. This is because Arm U has been revealed to be good and therefore first-order stochastically dominates Arm F. I refer to these two theoretical restrictions as the single-switching hypothesis.

Intuitively, because learning about Arm U is only possible while it is played and the option value of this information declines over time, optimal exploration is front-loaded. The decision maker therefore switches to Arm F once the posterior belief about Arm U's quality falls below a belief-dependent cutoff. This holds with a finite horizon, no discounting and any degree of constant absolute risk aversion over total payoffs. The formal result is provided in the Appendix 7.8.

H2 Risk-Neutral Exploration Benchmark. Given the experimental parameters, a risk-neutral Bayesian decision maker optimally explores Arm U for eight rounds in the absence of a reward before switching to Arm F (see Appendix 7.8 for derivation). Exploration durations below this cutoff constitute under-exploration relative to the risk-neutral benchmark. Exploration durations above this cutoff constitute over-exploration relative to the risk-neutral benchmark.

This cutoff therefore provides a quantitative benchmark for organizing exploration behavior with respect to the risk-neutral optimum. Unlike the optimal switching path, it imposes restrictions on preferences and cognition I do not observe, so it is a way to organize behavior more than an evaluation of it.

H3 Invariance to Switching Architecture. Under the optimal switching path, the optimal exploration strategy is invariant to restrictions on the number or timing of switching decisions, provided that the decision maker retains the opportunity to switch from Arm U to Arm F when the optimal cutoff is reached.

Given the structure of the problem, the optimal policy is characterized by backward induction. The decision maker can therefore determine the optimal switching cutoff at the beginning of the problem and, at any subsequent period, given the current history. Consequently, restricting when or how often the decision maker can switch does not alter the optimal exploration policy, provided the optimal switch remains feasible.

Together, these hypotheses characterize a sharp benchmark: the decision maker explores Arm U until either a reward is observed or a belief-dependent cutoff is reached, after which choices are permanent—Arm U in the former case and Arm F in the latter.

Deviations from this benchmark can arise through (i) returning to Arm U after switching to Arm F during the QU informational stage or (ii) any switch away from Arm U during the QC informational stage.

3 Experimental Design

Baseline Environment

In each problem, subjects chose between Jar U and Jar F[9] for 30 consecutive rounds. Each jar contained 10 balls. In every round, after a subject selected a jar, the computer randomly drew one ball from it, showed it on the screen, and returned it to the jar, so draws were independent across rounds.

Balls could be black or white. A black ball yielded 50 points, while a white ball yielded zero points. A subject's payoff for a problem was the sum of points accumulated across its 30 rounds traded at a rate of 1 USD per 100 points.

Jar F contained one black ball and nine white balls. This composition was known to subjects and fixed across all problems. Jar U could be of two equally likely types, determined at the beginning of each problem. A good Jar U contained two black balls and eight white balls, whereas a bad Jar U contained only white balls. A good Jar U was twice as likely to yield a reward as Jar F and therefore first-order stochastically dominated it, while Jar F first-order stochastically dominated a bad Jar U. Ex ante, before any draw from Jar U, both jars had the same expected probability of yielding a reward (10 percent). This is the simplest jar composition that aligns with the parameterization of the model used to test my hypotheses.

The experimental interface was designed to minimize ambiguity and maintain engage-

[9]In the experiment, the jars were named A/B instead of U/F.

ment. Subjects selected a jar by clicking on it, which triggered a brief animation in which a ball emerged from the top of the jar. The color of the ball, the associated payoff, and the updated cumulative score were displayed and remained visible while the program advanced to the next round. When a black ball was drawn from Jar U, the interface displayed a message indicating that Jar U had been revealed to be good. Subjects then clicked a "continue" button to proceed. This pause made the revelation of Jar U's type salient and marked the transition from the Quality Uncertain Stage—during which choices involved inference about Jar U's type—to the Quality Certain Stage, in which Jar U was known to be dominant.

To reduce cognitive frictions related to memory, attention, or misunderstanding of the environment, the interface displayed the round number, cumulative score, and number of unsuccessful attempts of Jar U. In addition, for Jar F, the screen displayed the probability of drawing a black ball. For Jar U, it displayed the initial probability that the jar was good[10] and the probability of drawing a black ball if Jar U were good. In the case that a reward was obtained from Jar U, the display updated to indicate that Jar U was good. Displaying the failures on Jar U and its revelation spares subjects from tracking their own history: in a bandit task, Banovetz and Oprea (2023) find that when the computer tracks past actions and outcomes for subjects, the share following the optimal exploration rule more than doubles. Figure 1 presents the instruction page explaining the information displayed in each round.

Overall, the environment is intentionally simple and transparent, allowing me to isolate exploration behavior from belief-updating errors, memory constraints, or ambiguity about payoffs.

Treatments

Subjects interacted with the baseline environment under several within-subject treatments that differed only in the restrictions imposed on switching behavior. Across treatments, the payoff structure, the information about the jars, and the stochastic process governing outcomes were held fixed.

Treatment 1: Free Switching. Treatment 1 corresponds to a canonical one-armed bandit with an arm whose payoff distribution is ex ante known and an arm whose payoff distribution is ex ante unknown. In each round, subjects were free to choose between Jar U and Jar F, with no restrictions on switching. They could move back and forth between jars arbitrarily.

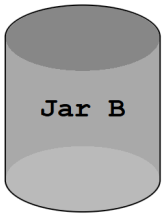
[10]To avoid confusion among subjects unfamiliar with Bayesian updating, I did not display posterior probabilities. A probability that changes mechanically over time could be misinterpreted as a change in the underlying environment rather than as belief updating.

Figure 1: Informational Page

Choice Screen Guide

When you make choices, the screen will remind you of how each jar was selected and how likely it is to give a reward if good. Below is an example:





The computer will show how many rounds you have chosen Jar A without drawing a black ball:

Draws without reward:

The message below tells you how the jar was selected and how likely it is to give a reward if good:
Reminder: Jar A was randomly chosen from a group with **50%** good jars.

A *good Jar A* returns a reward with probability **20%**.

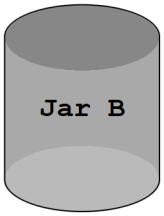
You don't know yet whether Jar A is good or bad.

The message below reminds you that Jar B is always good and how likely it is to give a reward.
Jar B is ***Good***.

It returns a reward with probability **10%**.

If you find a black ball from Jar A, you then know that the jar is good. The information message will update.
The picture below shows an example *after Jar A returned a black ball.*





The message below updates if Jar A is revealed to be good.
Update: You now know that Jar A is ***good***.

It returns a reward with probability **20%**.

The message below reminds you that Jar B is always good and how likely it is to give a reward.
Jar B is ***Good***.

It returns a reward with probability **10%**.

Next

This treatment serves as the benchmark environment, capturing unconstrained exploration behavior. Because choices were unrestricted, it allows a clean separation between behavior during the Quality Uncertain Stage—when the type of Jar U is unknown—and behavior during the Quality Certain Stage—when Jar U has been revealed to be good and is first-order stochastically dominant. In this environment, I observe subjects' initial jar choice, the total number of switches, and the frequency with which Jar U is chosen in each stage of the problem.

Treatment 2: Single Switching. Treatment 2 restricted switching behavior by allowing subjects to switch from Jar U to Jar F at most once. Once a subject selected Jar F, only Jar F remained available for the remainder of the problem. Because this restriction is irreversible, subjects were reminded of it when clicking on Jar F and were required to confirm their choice before proceeding. If Jar U was revealed to be good, only Jar U remained available for the remainder of the problem.

This constraint mechanically limits switching to the model prediction and allows exploration frequency to be measured independently of over-switching tendencies.

A limitation of this treatment is that the observation of exploration is constrained by the realization of a reward. If a reward from Jar U was realized early, the subject was forced into the Quality Certain Stage, and I do not observe how long the subject would have been willing to continue exploring in the absence of success. As a result, the frequency of choice of Jar U alone does not distinguish between short exploration driven by unfavorable outcomes and short exploration driven by early success. This censoring motivates the design of Treatment 3.

Treatment 3: Single Switching Ex Ante. Treatment 3 retained the single-switch restriction but introduced an ex ante planning component. Before starting each problem, subjects reported the number of rounds they intended to explore Jar U before switching to Jar F in the absence of a reward. Only one switch was permitted.

To familiarize subjects with this reporting mechanism, the first two problems in this treatment were non-binding. In these problems, subjects made an ex ante report but also played round by round, with a single switching opportunity as in *Treatment 2*. Payoffs for these problems were determined by their round-by-round choices. At the end of each problem, the interface displayed a comparison between the subject's ex ante report and the number of rounds actually spent exploring Jar U.[11]

[11] If a reward was obtained before the reported exploration length was reached, no comparison message was shown. If a reward was obtained after the reported exploration length, the message indicated that the subject explored beyond their ex ante report.

In the third and last problem, the ex ante reported exploration length was binding and enforced by the computer in the background (no round-by-round play at all). To push subjects to treat each ex ante report as a payoff relevant choice, I told them that only their report for the last problem would be enforced by the computer, but I did not tell them how many problems were in this treatment, so subjects could not know if a problem was the last one when they were making the ex ante report. After submitting the report, they were informed whether the current problem was the final one. In the final problem, subjects received a summary message indicating whether a reward was obtained from Jar U, the round in which it occurred (if applicable), and the resulting payoffs.

This treatment eliminates the censoring present in Treatment 2, since exploration length is chosen before any outcomes are realized. It therefore provides a cleaner measure of subjects' maximum willingness to explore Jar U. Moreover, by fixing exploration ex ante, it allows me to study exploration behavior in the absence of real-time reactions to unfavorable outcomes.

Treatment 4: Block Choice. This treatment serves a different purpose than Treatments 2 and 3. It is intended to evaluate the role of unexpected reactions to experience. Subjects do not choose jars round-by-round; instead, they start by simultaneously choosing an arm and the number of rounds they would like to automatically draw from this arm before choosing an arm again (the block size). In those rounds, they observe the outcome but cannot switch. When the rounds of the block are over, they choose an arm again (can switch or stay in the same arm) and a new block size for the new selected arm. The process continues until the problem ends.

The three key features of this treatment are as follows. First, the size of each block is chosen freely by the subject, so they can replicate the free-switching environment by choosing blocks of a single round. Second, the block size is chosen before any outcome is observed from the selected arm in that wave, eliminating responses to within-wave experience. Third, when choosing a block from Arm F, subjects can either commit to the rest of the problem, or retain the opportunity to reconsider at a posterior period.

Treatment 5: Ex Ante Rules. Rather than let subjects react to arm outcomes in real time, this treatment asks them to commit to a switching algorithm that the computer then executes mechanically in the background, showing them only the final score. Each problem splits into two decision points mirroring the Quality Uncertain/Quality Certain stages: First, subjects select one of four rules to govern switching for as long as Jar U has not yet paid off, and set that rule's parameter(s). Then, they choose a rule in case that Jar U returned a

reward: either let their initial rule keep running or override it and lock in Jar U for the rest of the problem.

The four rules include single switching, as in Treatment 3, and three other alternatives that permit multiple outcome- and non-outcome-driven switches between jars. Rules were presented in one of two possible randomized orders. Full details of each rule and presentation order are given in section 7.3 in the Appendix.

The goal of this treatment is to study switching choices in an environment in which round-by-round experience is fully muted, neither realized nor expected. Because treatments differ only in switching constraints and the timing of commitment, the optimal switching and exploration policies are identical across treatments.

Experiment Structure

I conducted three sessions for this paper. In Session 1, each subject completed 13 problems: five under Treatment 1, five under Treatment 2, and three under Treatment 3.

In Session 2, each subject completed 11 problems: five under Treatment 1, two under Treatment 2, three under Treatment 3, and one extra problem under Treatment 1.[12]

In Session 3, each subject completed 8 problems: five under Treatment 1, two under Treatment 4, and one under Treatment 5.

Within each session the order of treatments was fixed. This progression from unconstrained choice to environments with increasing structural restrictions allowed subjects to become familiar with the stochastic structure of the task before facing irreversible or planning-based constraints and facilitates interpretation of treatment differences as responses to progressively tighter switching restrictions.[13]

At the end of each session, one treatment and one problem within that treatment were randomly selected, and the cumulative payoff earned in that problem determined the subject's bonus compensation in addition to the participation fee. Prior to the experiment, subjects received detailed instructions and completed comprehension questions.

[12]Session 2 also included earlier versions of Treatments 4 and 5, which were revised for Session 3 (see its pre-registration); the analysis of these treatments uses Session 3 data only.

[13]Randomizing treatment order would require a substantially larger sample size and could confound treatment effects with adjustment costs or learning associated with relaxing switching constraints, which are not the focus of this study.

Experimental Procedures

The experiment was pre-registered in (AsPredicted). It was conducted online using the platform Prolific and implemented in oTree (Chen, Schonger, and Wickens, 2016). Participants received a participation fee of \$3 for session 1, \$4.5 for session 2, and \$3.5 for session 3. Across the three sessions, they received an additional \$0.01 for each point earned in the randomly selected problem. A total of 300 subjects participated in the experiment, each one allowed to participate in one session only. In total, 97 subjects participated in session 1, 98 subjects participated in session 2, and 105 subjects participated in session 3.

Before making their choices, participants completed a series of multiple-choice comprehension questions specific to each treatment. When a participant answered incorrectly, the computer informed them that one or more of their answers were wrong, but did not indicate which question(s). Typically, each page contained two to four questions. The number of attempts was recorded. For the initial instruction phase, the time to complete the comprehension questions was limited in order to maintain participants' attention, which is particularly important in an online setting. Participants who failed to complete the comprehension stage within the allotted time were not allowed to proceed to the choice phase and instead received only a partial participation payment. The complete set of instructions and comprehension questions is included in Appendix 7.12. For Treatment 4, participants additionally completed a practice problem before the payoff-relevant problems.

4 Results

I organize the results in five steps, following the model's hypotheses. First, I document deviations from the optimal switching path (H1) in both the QU and QC stages. Second, this behavior is associated with greater exploration and lower earnings. Third, I show that over-switching partially arises through reactions to uninformative outcomes, which utility curvature does not explain. Fourth, the Block Choice and Ex Ante Rules treatments show that over-switching reflects both execution and valuation distortions, and over-switchers classified accordingly differ in their free-switching behavior. Finally, imposing single switching reduces exploration among over-switchers who explore more than single-switchers, with little effect on single-switchers.

4.1 Switching Behavior and Consequences

Hypothesis 1: Switching Path

For Hypothesis 1 I use data from Treatment 1. In particular, for points (i) and (ii) I consider each subject's last problem in which the unknown arm (Arm U) never returned a reward[14], and for point (iii) each subject's last problem in which Arm U paid. I focus on subjects' last problem of each type to allow behavior to stabilize once subjects are familiar with the environment, limiting the influence of learning effects on the classification. In total, 295 subjects had at least one problem in which Arm U never paid and 267 had at least one problem in which Arm U paid.

(i) Starting arm. 222 subjects (75.3%) started with Arm U, following the model prediction. Of the 73 subjects who started with the familiar arm (Arm F), 36% did not explore Arm U at all, making them a minority of pure under-explorers. It follows that full unwillingness to explore was present in only 26 of the 295 individuals.

(ii) Returns after switching to Arm F in the Quality Uncertain informational stage. A total of 147 subjects (49.8%) returned to Arm U after having switched away to the Familiar one. The percentages are 52.3% for those who started with the predicted arm and 42.5% for those who started with Arm F. For this last case, since these subjects start with Arm F, I first require a switch to Arm U before considering it a model deviation. That is, I count a deviation when a subject first switches from Familiar to Unknown, then from Unknown to Familiar, and returns to Unknown at least once more. Among those who deviated, the mean number of switches is 7 for those who started with Unknown and 9 for those who started with Familiar. This also means that only 47.7% of correct starters followed the single-switching prediction, and even fewer, 21.9%, of incorrect starters followed a "quasi-single-switch" path, defined as a single switch to Arm U, or a switch to Arm U followed by a switch back to Arm F, since these paths mimic the correct switching path once I correct for starting with Arm F.

(iii) Changing arms after entering the Quality Certain informational stage. Of the 265 subjects who obtained a reward from Arm U before the last round, 38.9% switched away from it at least once after the breakthrough reward. However, behavior in the previous

[14] Although I analyze switching choices in the last problem in which Arm U never paid, the resulting classification captures meaningful differences in subjects' behavior throughout Treatment 1. Across the problems with at least five rounds in the QU stage, subjects classified as single-switchers in this problem have a higher share of problems with optimal switching, and a lower share with over-switching, throughout Treatment 1 (CDFs in Figure 5; Kolmogorov–Smirnov tests in Table 4 in the Appendix).

informational stage strongly predicts subsequent behavior. Among subjects who followed single switching before finding the reward, after-reward switching occurred for only 27.1% of the 166 subjects, with an average of 5 switches. This rate is more than double (67.7%), with an average of 7 switches, among the 62 subjects who over-switched before finding the reward.[15] Despite these switches, subjects stayed on Arm U for more than two-thirds of the remaining rounds after the first reward.

Single switching in both (ii) and (iii) is observed in 37.7% of the 260 subjects with both problem types. Deviation sizes across stages are significantly correlated ($\rho = 0.41$, $p < 0.001$), suggesting a classification of subjects into single-switchers in both stages (37.7%), over-switchers in both informational stages (31.9%), over-switchers who conform with the prediction in QC despite deviating in QU (21.5%), and others.

Result (Hypothesis 1: Switching Path).

- Although three-quarters of subjects start by exploring Arm U as predicted, roughly half subsequently return to it after switching away, and fewer than half of correct starters follow the single-switching path prescribed by the model.
- Switching deviations persist after Arm U is revealed to be good: even in the Quality Certain stage, where switching away is never informationally justified, 38.9% of subjects switch back to Arm F before returning to the now-dominant Arm U.
- Violations are not independent across stages: subjects who return to Arm U during the Quality Uncertain stage are substantially more likely to also switch away from it after Arm U is revealed to be good.

Robustness. The classification is robust to the problems used to classify switching behavior and to the quality of Arm U. Depending on the problems used, the share of subjects who single-switch at both stages ranges from 35% to 47.7% (Table 3 in the Appendix). Problems in which Arm U never paid can have a good or a bad Arm U, and because which problems these are depends partly on subjects' own sampling, single-switchers and over-switchers could have been classified on problems of different quality. However, the share of problems with a bad Arm U is similar for the two groups (Fisher exact $p = 0.108$), and the results hold when only problems with a bad Arm U are used (Appendix 7.5).

[15] 37 subjects who obtained a reward on the very first round of the problem, and 2 who obtained it on the last, were excluded from this analysis, since they have no meaningful pre-reward or post-reward segment.

Hypothesis 2: Risk-Neutral Exploration Benchmark

As in the analysis of Hypothesis 1, I use each subject's last problem in which Arm U never paid. I restrict attention here to the 269 subjects who explored Arm U at least once before switching, excluding the 26 subjects who never explored it. Within this sample, I distinguish *single-switchers*, who followed the single-switching prediction from H1 (or, for subjects who started with Arm F, its quasi-single-switch analogue), from *over-switchers*, who violated it by returning to Arm U at least once after switching away. Recall that the risk-neutral Bayesian benchmark prescribes exploring Arm U for eight rounds before switching and that exploration counts all choices of Arm U in the QU stage.

Exploration levels and switching behavior. Single-switchers explore Arm U for 9.27 rounds on average (median 9), closely tracking the eight-round benchmark. Over-switchers explore substantially longer, averaging 12.93 rounds (median 13). This gap is present and significant regardless of starting arm: among subjects who started with Arm U, over-switchers explore 3.68 rounds more than single-switchers on average ($p < 0.001$); among subjects who started with Arm F, the gap is 4.16 rounds ($p = 0.011$).

Differences in distribution between single- and over-switchers. I use single-switchers' exploration as the reference range. In the last Treatment 1 problem in which Arm U never paid, single-switchers ($n = 122$) explore Arm U for a median of 9 rounds, with an interquartile range of 6.25 to 10 rounds. Most over-switchers ($n = 147$) explore beyond this range: 68.71% explore more than the single-switchers' 75th percentile, 23.81% fall within the interquartile range, and only 7.48% fall below the 25th percentile. By construction, about half of single-switchers (52.46%) fall within this range. The pattern holds by starting arm. Among subjects who started with Arm U, 69.83% of over-switchers lie above the range, 25.86% within it, and 4.31% below it ($n = 116$). Among those who started with Arm F, the shares are 64.52%, 16.13%, and 19.35% ($n = 31$). A two-sample Kolmogorov–Smirnov test rejects that the two exploration distributions are equal ($D = 0.47$, $p < 0.001$), with the largest gap between the cumulative distributions at 10 rounds, the single-switchers' 75th percentile (Figure 2).

First Switch Away from Arm U. In their last problem in which Arm U never paid, over-switchers' first Arm U wave lasts 4.52 rounds on average (median 4), compared with 9.27 rounds for single-switchers, and this difference holds regardless of starting arm; 81.6% of over-switchers leave Arm U before reaching the eight-round benchmark. Most of their exploration therefore happens after they first return: on average, about 8.4 of their 12.93

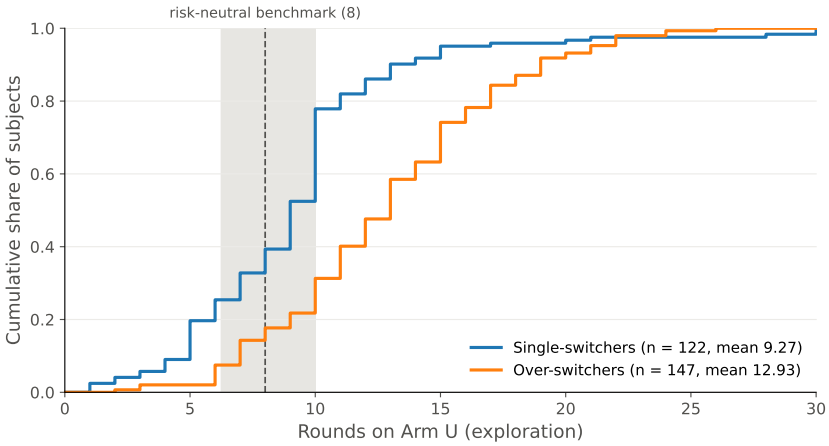


Figure 2: Exploration by Switching Group

Note: Cumulative distribution of rounds spent on Arm U (exploration) in each subject's last Treatment 1 problem in which Arm U never paid, for single-switchers ($n = 122$) and over-switchers ($n = 147$). The 26 subjects who never explored Arm U are excluded. The shaded band is the single-switchers' interquartile range (25th percentile: 6.25 rounds; 75th percentile: 10 rounds). The dashed line marks the risk-neutral benchmark of eight rounds. Two-sample Kolmogorov–Smirnov test: $D = 0.47$, $p < 0.001$.

rounds on Arm U, roughly two-thirds, come in later waves. These later waves show both margins: 50.3% of over-switchers leave Arm U again before their cumulative exploration reaches eight rounds, while 67.3% start at least one wave after having already reached it.

Premature abandonment therefore shows that bringing behavior closer to the model requires attention not only to unjustified returns to an option, but also to switching away before sufficient exploration has taken place. More broadly, the finding highlights a connection between the two components of the switching decision: whether a switch is justified by the information available and how much exploration takes place before that switch occurs.

Ruling out a mechanical explanation. Because every switch back to Arm U mechanically requires at least one more draw from it, excess exploration among over-switchers could in principle be mechanically explained just by switching more often. To rule this out, I compute, for each over-switcher, the minimum number of Arm U rounds consistent with their observed switch pattern (one per Arm U wave). Of the 121 over-switchers who explored more than eight rounds, 105 (86.8%) could still have stopped at the risk-neutral benchmark of eight rounds given their switch pattern. Of the 101 over-switchers who explored beyond the single-switchers' 75th percentile (10 rounds), 95 (94.1%) could still have stayed within it. Restricting the comparison to these mechanically feasible cases leaves the results essentially unchanged. Mean exploration among feasible over-switchers is 12.77 rounds under the eight-

round benchmark and 12.87 rounds under the 75th-percentile threshold, compared to 12.93 rounds unrestricted. The share exploring beyond the 75th percentile falls only marginally, from 68.71% to 67.38%. Excess exploration among over-switchers is therefore not an artifact of how switching is counted, but reflects a tendency to explore beyond what the switch pattern requires.

Result (Hypothesis 2: Risk-Neutral Exploration Benchmark).

- Deviations from the risk-neutral benchmark are not spread evenly across subjects: single-switchers explore close to the eight-round benchmark on average (9.27 rounds), while over-switchers explore about 3.7 rounds more (12.93 rounds), regardless of starting arm.
- Most over-switchers explore beyond the single-switchers' 75th percentile. The difference in distribution is significant.
- Deviations occur at both margins. Over-switchers abandon Arm U too early and return to an arm when the information does not justify it.
- This gap is not a mechanical consequence of switching more often: 94.1% of the over-switchers above the 75th percentile could have stayed within it given their switch pattern.

Welfare Consequences of Deviations from Single Switching

I now turn to the consequences of deviating from the single-switching policy for realized payoffs.

Problems in which Arm U never paid. Because Arm U never returns a reward in these problems, additional rounds exploring it displace rounds on Arm F, the only arm paying off ex post. Among the 269 subjects who explored Arm U at least once in their last problem in which Arm U never paid, single-switchers earn substantially more than over-switchers: 107.38 points ($1.07) on average, compared to 80.61 points ($0.81), a gap of 24.9% (Welch t-test $p = 0.003$; Mann–Whitney $p = 0.001$; Kolmogorov–Smirnov $p = 0.030$).[16] A regression of payoffs on the number of switches and starting arm confirms that each additional switch is associated with a 2.55-point ($0.026) reduction in payoff ($p < 0.001$), while starting arm

[16] I exclude the 26 subjects who never explored Arm U (see Appendix 7.4).

has no significant independent effect ($p = 0.639$). Full regression results are reported in Appendix Table 8.

Problems in which Arm U paid. In these problems, deviations from single switching are costly both before (QU) and after (QC) the first reward from Arm U. Among the 265 subjects who found a reward before the last round, QC-stage single-switchers earn 304.63 points ($3.05) on average, compared to 247.57 points ($2.48) for QC-stage over-switchers, a gap of roughly 20% ($p < 0.001$). To break this down by QU-stage behavior, I use each subject's last problem in which Arm U paid and exclude the 37 subjects paid by Arm U in the first round and the 2 paid in the last round, leaving 228 subjects classifiable at both stages. I count a return to Arm U after leaving it before the first reward as a QU-stage deviation. Subjects who follow single switching at both stages earn 309.50 points ($3.10) on average ($n = 121$); those who deviate only in the QC stage earn 258.89 points ($2.59, $n = 45$); those who deviate only in the QU stage earn 267.50 points ($2.68, $n = 20$); and those who deviate in both stages earn the least, 230.95 points ($2.31, $n = 42$). The 78.55-point ($0.79) gap between the extremes is significant ($p < 0.001$, Holm-adjusted).[17]

Part of this cost operates through delayed discovery of the good arm. Comparing subjects who obtain their first reward after the same number of Arm U draws, over-switchers reach it later: each Arm U draw takes them about half a round longer than single-switchers (0.8 rounds at the median), because they interleave draws from Arm F. For a reward that takes five Arm U draws, over-switchers obtained about four rounds later (Appendix Table 10). A regression of payoffs on switch counts in both stages separates the two costs. In the QU stage, the cost comes from delaying the first reward: each switch delays it by 1.36 rounds ($p < 0.001$) and is associated with 11.03 fewer points ($p < 0.001$), but only 0.67 fewer ($p = 0.77$) once the rounds remaining after the reward are held fixed. In the QC stage, the cost comes instead from leaving the good arm: 4.25 fewer points per switch with that control ($p = 0.028$), vs. 2.52 without it ($p = 0.188$). Full regression results are reported in Appendix Tables 9, 10 and 11.

Subjects were recruited on Prolific at a participation rate of about $9 per hour ($0.15 per minute). Had one of these problems been the one randomly drawn for payment, over-switchers would have lost about $0.27 in a problem in which Arm U never paid, which takes about 40 seconds, roughly 2.7 times this wage per minute of play. In a problem in which Arm U paid, which takes about one minute, they would have lost $0.57, about 4 times the

[17] Welch t-test p-values for the six pairwise comparisons, Holm-adjusted: 0.553 (both stages vs. QC only), 0.553 (both vs. QU only), < 0.001 (both vs. neither), 0.773 (QC only vs. QU only), 0.092 (QC only vs. neither), 0.442 (QU only vs. neither).

wage.[18] Subjects who deviate at both stages would have lost \$0.79 per problem in which Arm U paid relative to those who deviate at neither, about 5.2 times the wage.

Result (Welfare Consequences of Deviations from Single Switching).

- In problems in which Arm U never paid, over-switchers earn about 25% less than single-switchers, with each additional switch associated with lower payoffs, independently of the starting arm.
- In problems in which Arm U paid, subjects who deviate at both informational stages earn the least. The main cost arises from QU-stage switching: each switch before the first reward delays it by 1.36 rounds.

Robustness. The payoff cost of over-switching does not depend on the quality of Arm U, the problem used for classification, or luck. It holds when only problems with a bad Arm U are used (75.00 vs. 102.38 points, $p = 0.002$; Appendix 7.5). When the last problem (Problem 5), which all 300 subjects face, is used for classification, single-switchers earn 61.24 points more than over-switchers ($p < 0.001$), both when Arm U never pays (-34.89 points, $p = 0.003$) and when it pays (-44.10 points, $p = 0.095$; Appendix 7.4). Finally, over-switchers do not reach the first reward later because of bad luck: conditional on the number of prior unsuccessful draws, a draw from Arm U is as likely to pay for subjects who have already over-switched as for those who have not (difference +5.1 percentage points, $p = 0.12$; Appendix Table 11).

4.2 Mechanisms Behind Over-Switching

Experience and switching. Subjects may over-switch by responding to outcomes that, given the structure of the problem, carry no information about the relative quality of the two arms: the outcomes of Arm F, whose reward probability is known, and the outcomes of Arm U once it has paid and is known to be good. Table 1 summarizes the main results. All analyses use the five Treatment 1 problems and linear probability models in percentage points, with standard errors clustered by subject. Every comparison holds the subject and the position in the wave fixed; I indicate when the specification differs.[19]

[18] Payoff gaps are from the full sample; completion times are from Sessions 1 and 2, the only sessions where they were recorded.

[19] I use only the outcomes of rounds 1 to 28, because the experiment's code does not let me separate the outcomes of rounds 29 and 30 (Appendix 7.6).

Table 1: Responses to outcomes, over-switchers

	Coefficient	Std. error	p-value
Panel A. Arm F, QU stage (QU-stage over-switchers; 7,249 decisions, 147 subjects)			
(1) Arm F reward in the round just played	−0.40	2.08	0.849
(2) Rewards in the current Arm F wave	−3.10	1.00	0.002
(3a) Same, state held fixed	−2.79	1.42	0.050
(3b) Same, state held fixed, subject effects	−3.47	1.63	0.034
(4) Rewards in the current Arm F wave	−3.15	1.09	0.004
Rewards in earlier Arm F waves	+0.74	1.16	0.528
Difference	−3.88	1.56	0.013
Panel B. Arm U, QC stage (QC-stage over-switchers; row 5: 3,398 decisions, 102 subjects; row 6: 3,640 and 103)			
(5) Arm U reward in the round just played	−6.02	2.25	0.007
(6) Rewards in the current Arm U wave	−3.24	0.85	<0.001

Notes: Linear probability models; the dependent variable is switching away from the current arm in the next round. Coefficients in percentage points; standard errors clustered by subject. Sample: Treatment 1, all five problems, decisions that follow the outcomes of rounds 1 to 28 (the choices in rounds 2 to 29; footnote 19). Controls. Row 1: position in the wave interacted with the number of failures since the last reward in the wave, and subject fixed effects. Row 5: the same, with the position counted from Arm U's payment in the wave in which it paid. Row 2: position in the wave and subject fixed effects. Rows 3a–3b: one indicator for each combination of round, total earnings and failures on Arm U, and position in the wave; row 3a is the only row that does not hold the subject fixed. Row 4: position in the wave, failures on Arm U, round, and subject fixed effects. Row 6: rounds since Arm U paid within the wave, and subject fixed effects. Randomization $p \leq 0.001$ for rows 2 and 4 (current wave), and for row 6.

Arm F in the QU stage. I study the QU stage on the 147 QU-stage over-switchers. Once a subject is on Arm F, the outcome of a round is an independent random draw, so comparing a reward with a failure in the same round gives a causal effect. A first reward in the current Arm F wave has no detectable effect, but a second lowers the probability of switching away from Arm F in the next round by 10.4 percentage points ($p = 0.001$), and a third by 14.1 (Figure 3, top).

To separate a cumulative response to Arm F rewards (as if learning about a reward probability that is unknown but fixed) from a response confined to the current wave (a recency effect or attention limited to the current wave), I include the rewards in the current Arm F wave and those from earlier Arm F waves in the same regression.[20] Only current-wave rewards are associated with switching: each is associated with a 3.2 percentage point lower probability of switching, while rewards in earlier waves have no detectable effect (difference −3.9, $p = 0.013$; Figure 3, middle).[21] Each additional failure since the last reward, at the

[20] I also hold fixed the round and the failures on Arm U, so that the decisions compared have had the same number of Arm F draws, potentially differing only in their outcomes.

[21] A randomization test, which redraws every Arm F reward with its 10% probability and holds choices fixed, detects whether rewards have a nonzero effect (not its size) and is valid whatever the selection, because

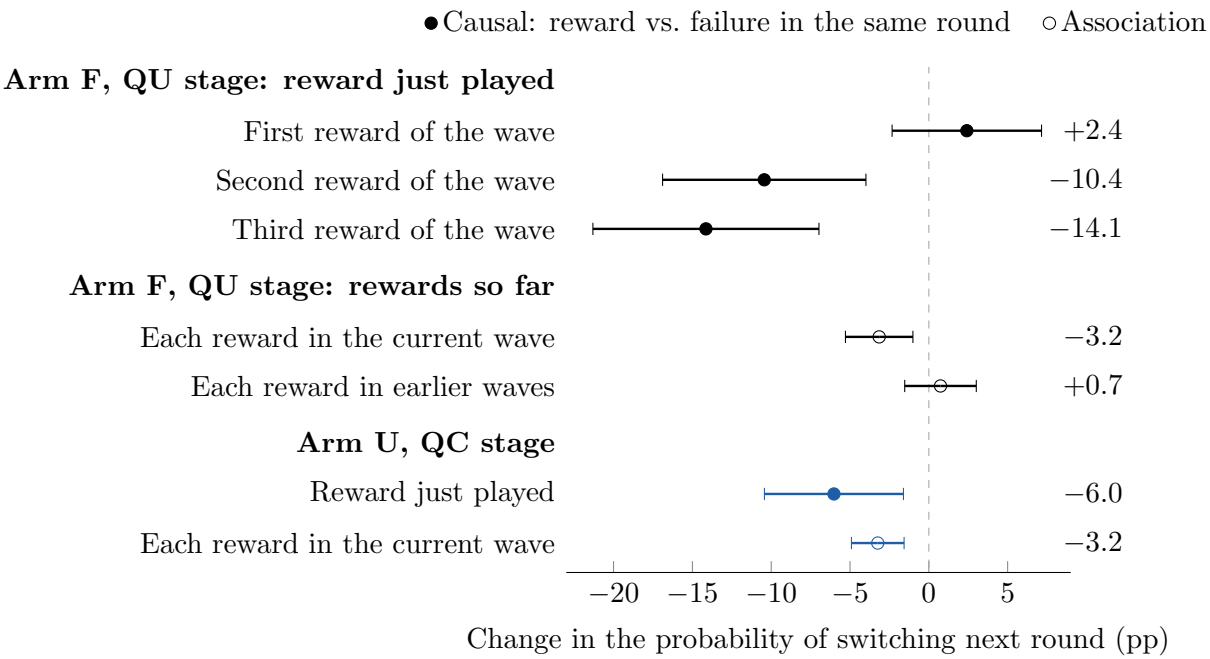


Figure 3: Responses to rewards that carry no information

Notes: Change in the probability of switching away from the current arm in the next round, in percentage points, with 95% confidence intervals (standard errors clustered by subject). Filled markers compare a reward with a failure in the same round and are causal; open markers are associations, because only waves that last long enough are observed (randomization tests in Appendix 7.6). Top: QU-stage over-switchers on Arm F, the reward just played by its order in the current wave (Panel C of Table 18; the 21 fourth or later rewards are omitted). Middle: rewards in the current and in earlier Arm F waves (row 4 of Table 1). Bottom: QC-stage over-switchers on Arm U (rows 5 and 6 of Table 1).

same position in the wave and with the same number of rewards, is associated with a 2.0 percentage point higher probability of switching away ($p = 0.007$; Appendix 7.6).

Arm U in the QU stage. Although every draw from Arm U is a failure before it pays, over-switchers rarely settle on Arm F: when Arm U was bad, they had stopped choosing it for good by round 25 in 30% of problems, against 71% for single-switchers (Table 20).

The QC stage. Once Arm U has paid, it is known to be good and first-order stochastically dominates Arm F. I study this stage on the 103 QC-stage over-switchers (Table 1, Panel B). A reward from Arm U in the round just played causally lowers the probability of switching away from it in the next round by 6.0 percentage points ($p = 0.007$; Figure 3, bottom; row 5 of Table 1), and each reward in the current Arm U wave is associated with a 3.2 percentage point lower probability of switching away from it ($p < 0.001$; randomization $p < 0.001$, see footnote 21; row 6 of Table 1).[22] At the same position in the wave and with the same number of rewards, each failure since the last Arm U reward is associated with a 2.7 percentage point higher probability of switching away from it ($p < 0.001$). Subjects who over-switch in both stages have a 6.6 percentage point lower probability of switching away from Arm U after a failure in the QC stage than in the QU stage ($p = 0.008$; 78 subjects), consistent with partial learning about the quality of Arm U.[23] On Arm F in this stage, the estimates for the next choice are too imprecise to be informative: with only 153 rewards, their standard errors are 2.6 to 4.1 times those of the QU stage (Table 21).

Alternative models. I next examine how standard alternative models of switching perform against these results (Table 22). The causal responses to uninformative outcomes, and the tendency of subjects to stay on the arm they are on,[24] rule out models in which switching is random (Panel B). Neither a constant switching rate (switching away from Arm F falls by 7.5 percentage points every ten rounds, $p < 0.001$) nor a rate that changes over time (randomization tests of footnote 21) holds in the data. Probability matching and Thompson sampling (Thompson, 1933) do not hold either, since they predict that, with the same information, the next choice does not depend on the arm currently chosen. Unlike subjects

if rewards did not affect choices, which waves last would not depend on the draws. Randomization $p \leq 0.001$ for rows 2 and 4 of Table 1 and for two or more rewards (Appendix 7.6).

[22] As in the QU stage, only rewards in the current wave are significantly associated with a lower probability of switching (Appendix 7.6).

[23] Appendix 7.6 reports the comparison over all decisions, not only failures, and within the same problem, where it is too imprecise to confirm the difference (-0.94 percentage points, $p = 0.87$).

[24] For the same subject, being on Arm U is associated with a 36.6 percentage point higher probability of choosing it next in the QU stage, at the same number of Arm U failures, and a 21.3 percentage point higher probability in the QC stage, at the same numbers of Arm U rewards and failures ($p < 0.001$ for both; Appendix 7.6).

in the matching literature, who deviate toward the more likely option, over-switchers deviate toward the option with the lower expected reward (Appendix 7.7).

Among the models in which choices respond to outcomes regardless of their informational content (Panel A), the lack of a detectable effect of a single Arm F reward on the next choice (-0.40 percentage points, $p = 0.849$; row 1 of Table 1) rules out win-stay/lose-shift, rules that switch after a fixed number of consecutive failures, and treating Arm F as of unknown quality, since a single reward would then reveal that Arm F, or the current wave, is good. That rewards in earlier waves have no detectable effect rules out learning about a fixed reward probability of Arm F and reinforcement learning in which its value carries over from one wave to the next.[25]

Taken together, a model that fits the data should treat switching as a non-deterministic decision that depends on the arm currently chosen, responds to uninformative outcomes only within the wave (in the QU stage, only as rewards or failures accumulate), and becomes less likely on Arm U once it has paid, consistent with learning.

The closest model is the hot-hand belief of Rabin and Vayanos (2010; see also Gilovich, Vallone and Tversky 1985; Rabin 2002), whose agents, like the subjects here, under-react to short streaks and over-react to long ones, weight outcomes by recency, and let beliefs fade. It cannot explain why the response stops at the wave, unless beliefs fade with time rather than with draws (footnote 25), or why a single reward lowers switching in the QC stage, where it predicts the opposite. It is also a model of beliefs, not of choices, and says nothing about the 88% of QU-stage over-switchers' switches that occur in waves without a reward (Appendix 7.6) or about the dependence of the next choice on the arm currently chosen.

Result (Mechanisms Behind Deviations from Single Switching).

- **Over-switchers respond to rewards that carry no information.** On Arm F, one reward does nothing, but rewards that pile up within a wave make switching less likely. Once Arm U is known to be good, its rewards still make switching less likely.
- **Only rewards in the current wave count.** Rewards from earlier waves on the

[25] Measuring the number of Arm F draws since a reward tests this model: only Arm F draws update its value, so current- and earlier-wave rewards the same number of draws back should count equally. They do not: the current-wave reward is associated with a 5.2 percentage point lower probability of switching ($p = 0.003$; Table 22), a more precise comparison than holding fixed the rounds since the reward, because recent earlier-wave rewards are more common in draws than in rounds.

same arm have no effect, so the response resets every time they switch. That rules out learning about the arm and reinforcement carried across waves.

- The data reject win-stay/lose-shift, threshold rules, random switching, probability matching and Thompson sampling. A hot-hand belief comes closest, but it can't explain why the effect resets every time they switch.

Robustness. Current-wave rewards are usually also more recent than earlier-wave rewards. Holding fixed the number of rounds since the reward occurred, the gap between them is 3.3 percentage points ($p = 0.100$), close to the 3.9 points of row 4 of Table 1 but imprecise, because recent rewards from earlier waves are rare; holding fixed the number of Arm F draws since the reward instead, the gap is 5.2 percentage points ($p = 0.003$; footnote 25). The estimates are robust to fixed-effects logit estimation instead of linear probability models, to conditioning on Arm U being bad, and to alternative definitions of over-switchers and other samples (Appendix 7.6).[26]

4.3 Three Hypotheses for Over-Switching

I now consider three hypotheses for over-switching. In the canonical model, three conditions are jointly sufficient for single switching: preferences display no wealth effects, as under constant absolute risk aversion (CARA); the subject values the arms as the model does; and the subject carries out the choice that this valuation prescribes. Each hypothesis relaxes one of these conditions. The first, utility curvature, can be assessed with Treatment 1 alone; the other two require additional treatments.

Utility Curvature

Under CARA, a reward raises earnings but leaves preferences over the remaining choices unchanged. Under other utilities over earnings, such as constant relative risk aversion or utilities built around an aspiration level, a reward, or a run of rounds without one, can change which arm is preferred, so curvature could in principle generate over-switching. Four results argue against this explanation (Appendix 7.9.1). First, curvature can only explain deviations in the QU stage: once Arm U has paid, it first-order stochastically dominates Arm F, so no increasing utility prefers to switch away from it. Yet the same response to rewards appears in that stage (Panel B of Table 1), mostly among the same subjects (83 of

[26] Conditioning on Arm U being bad ensures that its not paying does not depend on the subject's own choices. Appendix 7.6 also compares the estimates with subjects who are not over-switchers.

the 103 QC-stage over-switchers are also QU-stage over-switchers). Second, within a problem wealth can only rise, so a return to Arm U driven by wealth must follow a reward, as under constant relative risk aversion; switching should then rise after a reward rather than fall. Third, the response survives once I hold fixed everything a utility over earnings and the remaining horizon can respond to, aspiration levels included. With one indicator for each combination of round, total earnings and failures on Arm U, and therefore no functional form, current-wave rewards are still associated with a 2.8 percentage point lower probability of switching ($p = 0.050$), and 3.5 with subject fixed effects ($p = 0.034$; state held fixed, rows 3a–3b of Table 1), concentrated at two or more rewards (8.4 and 10.6 percentage points; $p = 0.010$ and $p = 0.004$). With total earnings fixed, this compares when rewards came, not how many. Fourth, standard utilities fitted to each subject's choices imply implausible risk attitudes, and no fitted utility, aspiration utilities included, reproduces the response to current-wave rewards when simulated (Table 24). Utility curvature is therefore not a plausible main driver of over-switching.

Execution and Valuation Distortions

The remaining two hypotheses relax the other two conditions, which in the formal framework amounts to adding one distortion to the model. Under the first, *noisy execution* of a correctly identified policy, subjects broadly recognize the single-switching strategy as optimal but deviate from it as they experience the problem. Under the second, *distorted valuation*, deviations reflect preferences or valuations that differ from the benchmark, such as an inherent preference for switching or an incorrect idea of how to maximize expected payoffs. Noisy execution includes random errors and reactions to experience or outcomes, such as frustration, second-guessing, or boredom. The key distinction is that execution-based deviations arise through experience, whereas valuation-based deviations persist even when the subject commits to a rule before experiencing any outcome. These hypotheses imply different responses to restrictions on experience. Because Treatment 1 alone cannot distinguish between them, I develop a formal framework in Appendix 7.9.2 and introduce additional treatments to generate testable predictions. Here I provide a conceptual description.

Execution-based distortions. Experience during play can affect the propensity to stay on or switch from an arm beyond the belief-updating mechanism of the canonical model. I model these distortions at the wave level: they depend on consecutive pulls from the same arm, reset when the subject switches arms, and play the role of an extra term added to the standard valuation of the arm; this term distorts the subject's assessment of the current arm

relative to the other, from the perspective of the current period. For example, unsuccessful draws from Arm F may create an additional pull toward switching for some periods to Arm U, even though they provide no new information about the state of the problem.

Some subjects may anticipate that within-wave experience could make them change their mind, and so value the flexibility to react to it; I then say that their execution distortions come with a *demand for flexibility*, in a sense similar to Kreps (1979).

Valuation-based distortions. The subject's underlying valuation of the arms differs from the canonical benchmark. Such distortions are stable across waves, persist when the subject switches arms, and can alter the relative value assigned to exploration and exploitation. A valuation distortion can affect the QU stage (before Arm U pays), the QC stage (after it pays), or both, so I treat the two stages separately. Valuation-based and execution-based distortions are not mutually exclusive: a subject may value the arms differently from the benchmark and also execute their strategy with within-wave noise.

To isolate the source of the deviations described above, I implement two alternative *choice architectures* — protocols that change when and how often subjects make decisions, and what they have observed at each decision point.

Definition: Block Choice. Under Block Choice, whenever the agent chooses an arm, they also choose a number of periods to automatically draw from that arm, forgoing the opportunity to revise the arm during that block. Thus, the agent chooses sequences of same-arm draws rather than an arm in every round. If Arm U returns its first reward, the block restarts as the informational stage changes. I implement this architecture in Treatment 4 (Block Choice).

Under Block Choice, a subject sets the length of the first block, and of the block that follows Arm U's first reward, before any within-wave experience can act. Hence, a subject affected only by execution distortions without a demand for flexibility covers their whole planned exploration of Arm U in the first block and their whole stay on Arm U in the block after it pays, and, once they switch to Arm F, stays there to the end: they follow a single-switching path (see Proposition 6 in Appendix 7.9.2), despite every path being available.[27] Execution distortions with a demand for flexibility may persist: such a subject chooses shorter blocks, and at the end of each block they include the current distortion, if present, in their choice. Valuation distortions also persist, as they are present throughout the problem.

Definition: Non-Experienced Choice. A choice architecture is called non-experienced

[27]Single switching need not occur at the optimal cutoff: experience on Arm U may lead the subject to extend exploration when the first block ends.

if the agent commits, before the problem starts, to a complete contingent rule for the entire problem, and observes no outcomes until it ends. I implement a proxy of this choice architecture in Treatment 5, Ex Ante Rules.

Because subjects observe no within-wave outcomes, execution distortions have no experience on which to operate and nothing to anticipate. Thus, subjects affected only by execution distortions, with or without a demand for flexibility, follow a single-switching path under Non-Experienced Choice (see Proposition 7 in Appendix 7.9.2). Valuation distortions, however, remain, so subjects with them still do not choose a single-switching rule.

Block Choice: Results As pre-registered, both treatments target over-switching, so I restrict the analysis to subjects who over-switched in Treatment 1.

In the QU stage, half (16 of 32) of the subjects who over-switched in Treatment 1 (problem in which Arm U never paid) become single-switchers under Block Choice;[28] 13 continue to over-switch and 3 fail to explore. The share is somewhat higher in the QC stage: 13 of 21 Treatment 1 QC over-switchers (62%) become single-switchers, while 6 continue to over-switch and 2 switch away from Arm U as soon as it pays and never return.

Block Choice also increases initial exploration: over-switchers' mean first Arm U wave rises from 3.76 rounds in their last Treatment 1 problem in which Arm U never paid to 10.02 in their first Block Choice problem ($p < 0.001$, $n = 41$; Appendix 7.9.3).

Ex Ante Rules: Results Ex Ante Rules implement a proxy of Non-Experienced Choice: subjects choose from a restricted menu of rules, all starting with Arm U and including the single-switching path, so a subject whose plan is not single-switching chooses the closest available rule.[29]

Among the 31 subjects who over-switched in Treatment 1's QU stage and also faced Ex Ante Rules,[30] 17 chose a single-switching rule under Ex Ante Rules,[31] compared with 15 under Treatment 4 (McNemar exact $p = 0.77$). I use only whether the chosen rule is single-switching, not its form, and I do not test the starting arm.

[28] Two of these 16 start with Arm F and follow a quasi-single-switch path, and two start with Arm U and never switch away from it. The QU classification requires a Block Choice problem in which Arm U never paid, which 32 of the 42 Treatment 1 QU over-switchers who played Block Choice had; in the QC stage, 21 of the 35 had one in which Arm U paid.

[29] The share choosing a single-switching rule is therefore an upper bound.

[30] Of the 32 subjects classified under Block Choice, one did not complete Ex Ante Rules because Prolific timed them out.

[31] One of them did not choose the single-switching rule but a failure-based rule that allows only one switch.

For the QC stage, among the 21 subjects who over-switched in Treatment 1, 12 single-switch under Ex Ante Rules, against 13 under Treatment 4: they lock in Arm U once it pays. The remaining 9 keep a cycle or a rule that switches away from Arm U after it pays. Ten subjects single-switch under both treatments.

Combining the Two Treatments Neither of these two hypotheses alone fits the data. If over-switching came only from execution distortions, almost all over-switchers would have chosen a single-switching rule (Proposition 7); 45% did not in the QU stage, and 43% in the QC stage. If it came only from valuation distortions, almost none would have (Proposition 8(a)); instead, 55% in the QU stage and 57% in the QC stage chose a single-switching rule, a share too large to come mostly from subjects pushed to the single-switching rule by the restricted menu, which offered several rules that switch more than once. Over-switching therefore reflects both sources.

Table 2 shows how the two sources divide across subjects.[32] Those who single-switch under both treatments are consistent with execution distortions, and those who single-switch under neither with a valuation distortion. In the QU stage, the five subjects who single-switch only under Ex Ante Rules can be rationalized by execution distortions with a demand for flexibility; the remaining cases are discussed in Appendix 7.9.3.

Table 2: Over-switchers by ex ante rule and Block Choice behavior

Panel A: QU stage	Single-switch	Over-switch	Never choose Arm U
Single-switching ex ante rule (17)	10	5	2
Other ex ante rule (14)	5	8	1
Panel B: QC stage	Stay on Arm U	Over-switch	Switch away from Arm U at once
Lock in Arm U (12)	10	1	1
Other ex ante rule (9)	3	5	1

Notes: Treatment 1 over-switchers in each stage who also faced Ex Ante Rules; Block Choice behavior in the classifying problem (one in which Arm U never paid for the QU stage, one in which it paid for the QC stage). In Panel A, "Single-switch" includes two paths that start on Arm F. In Panel B, "Stay on Arm U" means staying on Arm U from its first reward to the end of the problem, in one block or several.

[32] Propositions 6 and 8 in the Appendix formalize this classification.

Result (Classification Framework: Execution-Based versus Valuation-Based Distortions).

- Block Choice moves roughly half of QU over-switchers and 62% of QC over-switchers to single-switching, while increasing initial exploration of Arm U.
- Ex Ante Rules move about as many over-switchers to single switching as Block Choice: 17 versus 15 of 31 in the QU stage, and 12 versus 13 of 21 in the QC stage; under the framework, the over-switching of those who single-switch under Ex Ante Rules is consistent with execution distortions in that stage.
- Neither source alone explains over-switching: 45% of QU-stage and 43% of QC-stage over-switchers do not choose a single-switching rule, which execution distortions alone cannot produce, while 55% and 57% do, which valuation distortions alone cannot produce.
- Crossing the two treatments, 10 QU-stage over-switchers single-switch under both and 8 under neither, and 5 single-switch only under each of them; in the QC stage, the counts are 10, 5, 1 and 3.

Robustness. Subjects who single-switch under either treatment over-switched in 79% to 89% of their other Treatment 1 problems, so their change is not regression to the mean. In the QU stage, they had fewer extra switches in Treatment 1 than subjects who keep over-switching, with no trend over their last Treatment 1 problems; in the QC stage, some learning cannot be ruled out (Appendix 7.9.4). Both treatments always follow Treatment 1, and Ex Ante Rules always follow Block Choice.

4.4 Beyond Choice Restrictions: Behavioral Signatures of Over-Switching

Treatments 4 and 5 sort over-switchers by how they respond to each architecture. The strongest treatment responses are those that agree across the two. In the QU stage, 10 over-switchers single-switch under both Block Choice and Ex Ante Rules, consistent with noisy execution, and 8 keep over-switching under both, consistent with a valuation distortion. In the QC stage, 10 stay on Arm U under both, consistent with noisy execution in that stage,

and 5 keep switching under both, consistent with a valuation distortion in that stage.[33] I use these subjects as anchors and ask two questions. First, do the anchors differ in their free switching in Treatment 1, so that their response to the treatments could have been predicted from it? Second, do these differences characterize the other Treatment 1 over-switchers?

Free switching markers per anchor group. In the QU stage, the anchors differ in free switching; in the QC stage, they do not. The QU-stage anchors differ on 9 of the 21 measures I consider, most of them related to how often they switch and how they react to failures (Table 30 in Appendix 7.10). As a comparison, only 3.4% of 10,000 random splits of the Treatment 1 over-switchers into groups of 10 and 8 differ on as many measures. I focus on the three measures closest to my analysis of Treatment 1. Valuation-distortion anchors switch more often (switch rate 0.35 vs. 0.18; permutation $p = 0.026$), switch more readily right after a failure (0.40 vs. 0.21; $p = 0.028$), and follow the optimal single-switching path in fewer problems (0.18 vs. 0.50; $p < 0.001$).[34] These differences predict how each anchor responds to the treatments: assigning each anchor to the closer of the two group centroids in the space of the three markers, with the centroids computed without it, places 16 of the 18 in their own group, which random relabelings of the anchors achieve in only 0.1% of cases.

The 10 and 5 QC-stage anchors do not differ significantly in any of 19 measures of their QC-stage behavior in Treatment 1 (Table 31 in Appendix 7.10). This is despite their responses to the treatments being consistent: 15 of the 19 QC-stage over-switchers whose Block Choice behavior can be classified respond in the same way to both treatments, against 53% expected if the two responses were unrelated (Fisher exact $p = 0.041$).

Extending the classification. I assign each of the 147 Treatment 1 over-switchers to the QU-stage anchor centroid closest in Euclidean distance, in the space of the three markers above, standardized on the full sample (Figure 4). This yields 84 noisy-type and 63 valuation-type subjects (types, because their label is extrapolated from the anchors rather than observed under the treatments). The classification broadly agrees with a clustering that does not use the anchors and chooses its own variables (87.8% of labels recovered, Cohen's $\kappa = 0.74$), reflects stable traits of subjects (intraclass correlations of switch rate and failure-reactivity across problems: 0.67 and 0.64), and holds out of sample (78.3% of labels

[33] Appendix 7.10 reports the other over-switchers who played both treatments and how those in the QU stage are classified.

[34] Switch rate and failure-reactivity are measured in the subject's last Treatment 1 problem in which Arm U never paid; the share of optimal problems uses all five Treatment 1 problems.

reproduced in a separate Treatment 1 problem, $\kappa = 0.56$).[35] The QC-stage anchors cannot be told apart in free switching, so they give no classification to extend.

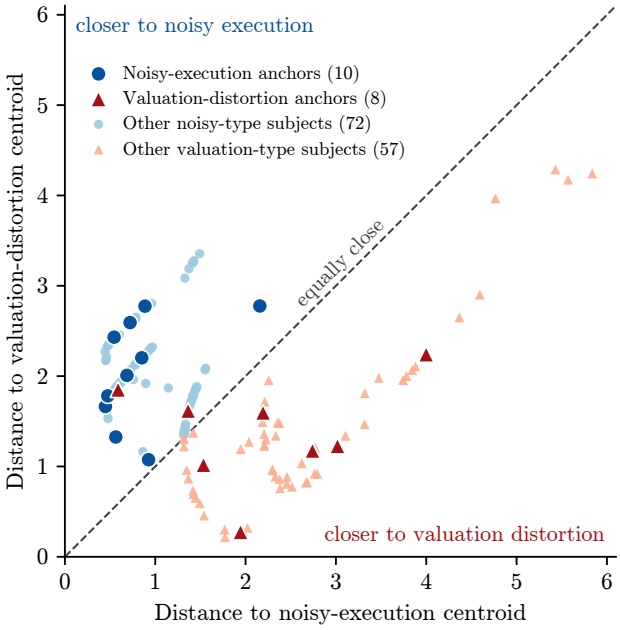


Figure 4: Classification of Treatment 1 over-switchers

Notes: Each marker is one of the 147 Treatment 1 over-switchers, placed by its Euclidean distance to the two QU-stage anchor centroids in the space of three standardized markers: switch rate, switching right after a failure, and the share of Treatment 1 problems with optimal single switching. Subjects above the dashed line are closer to the noisy-execution centroid and are classified as noisy type; those below it, as valuation type. Dark markers are the anchors: 10 over-switchers who single-switch under both Block Choice and Ex Ante Rules, and 8 who keep over-switching under both. Two valuation-distortion anchors lie closer to the noisy-execution centroid. Overlapping anchors are shifted slightly along the dashed line, without changing sides.

However, the QU-stage classification also predicts behavior in the QC stage. Among the 141 classified subjects with a problem in which Arm U paid, 79% of valuation-type subjects over-switch in the QC stage of their last problem in which Arm U paid, against 43% of noisy-type subjects (Fisher exact $p < 0.001$), and they make more extra switches (6.5 vs. 1.8; Mann–Whitney $p < 0.001$). QC-stage over-switching is therefore more common among valuation-type subjects.

[35] In a free-switching problem played after the restricted treatments, the valuation-type group still switches more (0.25 vs. 0.09) and reacts more to failures (0.30 vs. 0.11; $p < 0.001$ for both). Appendix 7.10 reports the details and further checks.

Are the two groups best predicted by different models? If the two groups reflect different sources of over-switching, their switching should also be organized differently. As I define them, noisy execution has a single-switching plan as its baseline (stay on Arm U until a personal number of failures, then move to Arm F for good), from which noise makes the subject depart at times; a valuation distortion does not. I test this with a horse race between two families of models, estimated subject by subject. *Plan models* give each subject a cutoff on Arm U failures, which may vary around a personal center from one problem to the next, and let them deviate from the plan either at random or more often as failures or time on an arm accumulate. *Models without a plan* give each subject a baseline propensity to switch that responds to failures in the current wave, to the length of the wave, or to their belief that Arm U is good.[36] The main parameters (baselines, slips, cutoffs and slopes) are estimated at the individual level, from problems other than the one being predicted, so all comparisons are out of sample.[37] I measure accuracy as the average (geometric mean) probability that a model assigns to the subject's actual choices.

In the QU stage, a plan predicts noisy-type subjects better than any model without one: the group's best plan model reaches an accuracy of 66.6%, against 61.0% for the group's best model without a plan (a difference of 5.6 points, bootstrap $p < 0.001$), and a plan model is the best model for 61% of these subjects.[38] For valuation-type subjects the two families cannot be told apart (52.5% vs. 50.9%; $p = 0.29$), and a plan is the best model only for 44% ($p = 0.033$ for the difference between the two shares). The advantage of the plan is larger for noisy-type subjects (rank test on each subject's out-of-sample gain from the plan model, $p = 0.002$), and the anchors, whose labels do not depend on Treatment 1, split the same way: a plan is the best model for 8 of 9 noisy-execution but only 3 of 8 valuation-distortion anchors ($p = 0.043$). When I cap the slip probability (switching when the plan says to stay) so that slips are rare, as true slips should be, the plan remains the best model for about as many noisy-type subjects (61% to 64%), while for valuation-type subjects this share falls from 44% to 25% and, under the tightest cap, models without a plan predict them significantly better (Appendix 7.11).

[36] I do not include a constant propensity to switch alone, since Section 4.2 rejects purely random switching. Appendix 7.11 describes the models and their estimation.

[37] Both families let the average number of switches differ by problem through the same problem fixed effects. These and the spread of the cutoff in plan models are estimated separately for each group. The results hold when these group-level parameters, and the group labels, are estimated without the predicted problems (Appendix 7.11).

[38] In the QC stage, the two groups (18 and 15 subjects) are too small for the horse race to separate plan models from models without a plan (Appendix 7.11).

This predictive exercise adds to the characterization of noisy-type and valuation-type subjects. Across all subjects, the relative predictive advantage of the plan model falls as subjects get relatively closer to the valuation-distortion centroid (Spearman correlation -0.26, $p = 0.002$). This is not a difference in how much subjects switch: holding each subject's switching rate fixed, a subject's position still predicts the gain from the plan model ($p = 0.013$), while the switching rate adds nothing ($p = 0.27$).

Result (Beyond Choice Restrictions: Behavioral Signatures of Over-Switching).

- Over-switchers who respond differently to the choice-architecture treatments differ in their free switching in the QU stage, and their response can be predicted from it; random groups rarely differ as much. In the QC stage, they respond consistently across the treatments, but their free switching does not tell them apart.
- The QU-stage classification extends to all 147 Treatment 1 over-switchers: it is stable across problems, recovered without the anchors, and persists after the restricted treatments. Valuation-type subjects are also more likely to over-switch in the QC stage.
- Out of sample, a single-switching plan predicts noisy-type subjects better than models without a plan, even when slips (switching when the plan says to stay) must be rare. Its advantage falls as subjects get relatively closer to the valuation-distortion centroid.

Robustness. This analysis was not pre-registered. Its results are robust to the choice of anchors and markers: the labels of the 147 subjects barely change when one anchor is left out of the centroids (2.1% on average) or the anchors are resampled (Appendix 7.10); and the results hold when the extrapolation and the horse race use all 9 measures on which the anchors differ (Appendix 7.10). Finally, valuation-type subjects over-switch more in the QC stage (79% vs. 43%) not only because they switch more in the QU stage: the difference remains when the QU-stage switch rate is held fixed (logit, $p = 0.012$), and when the problem in which QC-stage over-switching is measured is left out of the classification (78% vs. 37%; $p < 0.001$; Appendix 7.10).

4.5 Imposing Single Switching

Hypothesis 3: Switching Constraints

Hypothesis 2 documented a close relationship between over-exploration and informationally unjustified switches in the QU stage. Constraining switches provides a diagnostic environment to further understand how these behaviors interact.

If exploration were independent of switching dynamics, exploration patterns would remain unchanged when switching opportunities are constrained. Instead, over-switchers who explored more than single-switchers in the baseline treatment reduce exploration when switching is limited, while single-switchers remain stable.

I compare the last Treatment 1 problem in which Arm U never paid with the last Treatment 2 problem in which Arm U never paid and the last ex ante report of Treatment 3 (the one that is executed by the computer). Among subjects who over-switch and explored beyond the single-switchers' 75th percentile in Treatment 1 (10 rounds), mean exploration falls from 15.2 rounds in Treatment 1 (median 15) to 12.2 rounds in Treatment 2 (median 11.5) and 12.2 rounds in Treatment 3 (median 10). Both reductions relative to Treatment 1 are statistically significant. For the general over-switching group, the changes are smaller—2.0 rounds in Treatment 2 and 2.3 rounds in Treatment 3—and only the latter difference is significant at the 1% level.

These differences extend to the distribution. Among over-switchers, the CDF of Arm U exploration shifts toward fewer rounds under both constrained treatments, and Kolmogorov–Smirnov tests reject equality with the free-switching distribution. For single-switchers, I find no significant evidence of a distributional difference. I also find no significant difference between Treatments 2 and 3 in either means or distributions.

Among the 66 subjects who over-switch and explore beyond the single-switchers' 75th percentile, 71% reduce exploration in Treatment 2 and 74% in Treatment 3. Of those who reduce exploration, 26% move into the single-switchers' interquartile range in Treatment 2 and 43% do so in Treatment 3; 32% and 27%, respectively, fall below its 25th percentile. The remainder reduce exploration but remain above the 75th percentile. The CDFs and tests for over-switchers are reported in Figure 6 and Table 5 in the Appendix.

Single-switchers show no significant differences in mean exploration or its distribution across treatments. Among those who explored within their interquartile range in Treatment 1, 62.9% stay within it in both restrictive treatments. The CDFs and tests for single-switchers are reported in Figure 7 and Table 5 in the Appendix.

Result (Hypothesis 3: Switching Constraints)

- Over-switchers who explored more than single-switchers in Treatment 1 reduce exploration under both constrained treatments. This reduction is significant in both Treatment 2 and Treatment 3.
- Single-switchers' exploration does not change significantly under switching constraints.

Robustness. An alternative way to measure the treatment changes is to compare them with each subject's mean over all their Treatment 1 problems in which Arm U never paid. Against this mean, the reduction for over-switchers who explored more than 10 rounds is slightly smaller but still significant: 2.7 rounds in both Treatment 2 ($p = 0.009$) and Treatment 3 ($p = 0.005$), against 3.0 rounds in the main comparison ($p = 0.004$ and $p = 0.003$) (Table 6 in the Appendix). Once switching is free again, in Session 2's additional free-switching problem played after Treatment 3, the gap returns: over-switchers again explore more than single-switchers (13.4 vs. 8.8 rounds, $p < 0.01$), so the reduction comes from the constraints rather than from learning.[39] Across all single-switchers, 61.5% stay in the same part of the distribution (below, within or above the interquartile range) in all three treatments, and the share within the interquartile range barely moves (53.8%, 50.8% and 52.3% in Treatments 1, 2 and 3; Table 7 in the Appendix).[40]

5 Discussion

The results show that switching does not track information for a large share of subjects, that this persists once nothing is left to learn, and that it has two separable sources. Among subjects who faced both stages, about two in five follow single switching in both, about one in three over-switch in both, and about one in five over-switch only while quality is uncertain. This section discusses what this implies for bandit tasks as measurement instruments, how the results relate to probability matching and the description–experience gap, what they imply for the design of choice procedures, and an open question for strategic experimentation.

[39] Exploration under Treatment 2 differs between sessions (15.0 rounds in Session 1 vs. 8.3 in Session 2), possibly because Session 2 had 2 Treatment 2 problems instead of 5 (Appendix 7.2).

[40] The pattern is similar for subjects who single-switch in both stages: 61.3% of those who explored within the interquartile range in Treatment 1 stay within it in both restrictive treatments, and 56.6% stay in the same part of the distribution in all three.

What Do Bandit Tasks Measure? The canonical one-armed bandit is a useful device for studying willingness to explore, risk preferences, and responses to changes in the environment (e.g., Banks, Olson, and Porter, 1997; Hudja and Woods, 2024a), because it reduces sequential choice to a single trade-off between the option value of information and its cost. That reduction relies on a primitive that holds for some subjects but not for others: that subjects respond to an outcome only insofar as it carries information about the relative quality of the alternatives. About half of subjects return to the unknown arm (Arm U) without new information about it, and many abandon it after learning that it is good. The posterior belief induced by the history is therefore not sufficient to predict switching: a subject may abandon an arm and later return to it with no information about it in between, which no model in which beliefs about an arm change only with its own outcomes can justify. A model that reduces history to a posterior belief will misclassify part of what it measures.

This also affects how bandit experiments measure exploration itself. Hudja and Woods (2024a), who impose single switching, find that subjects explore less than the model predicts and that risk aversion does not account for the gap. My results show that measured exploration also depends on how the switching decision is structured: over-switchers who play Block Choice start with a first Arm U wave of 10.02 rounds on average, against 3.76 in their last Treatment 1 problem in which Arm U never paid, with no change in the learning problem. Restricting subjects to a single switch also reduces exploration among the over-switchers who explore most. If the same subject can look more or less willing to explore depending on how switching decisions are made, deviations from the benchmark do not only reveal a taste for information; they also reflect how subjects execute their plans and, for some, an incorrect model of the problem or non-standard preferences.

Separating these channels is harder in environments where no arm's quality is ever revealed, such as those of Gans, Knox, and Croson (2007) and Hudja and Woods (2025). There, every outcome is informative, which makes these environments well suited to asking which decision rule best organizes behavior, but not to the question I ask here: separating a switch driven by information from one that is not requires outcomes that carry no information. My design provides them in two ways: the familiar arm's outcomes vary but carry no information, and the unknown arm's quality becomes fully known once it pays. This lets me identify departures from the prediction with and without quality uncertainty, without assumptions about how beliefs are formed, at the cost of studying a simpler environment than many applications of interest.

Relation to Probability Matching and the Description–Experience Gap. Once Arm U pays,

the bandit becomes a repeated choice between two options with known reward probabilities, the setting in which the probability-matching literature finds that people choose each option in proportion to its probability of paying off (Vulkan, 2000). Compared with studies that state the probabilities, subjects in this stage behave much like the subjects of those studies and, like them, move toward the better option with experience (Appendix 7.7). QC-stage over-switchers, however, do not match. On average they choose Arm U about as often as matching implies, but few are individually close to the matching share, and their share of Arm U falls within a problem after Arm U pays. Their switching also responds to rewards in the current wave and resets when they switch, which probability matching does not predict. Before Arm U pays, over-switchers deviate toward the option with the lower expected reward, whereas in that literature deviations go toward the more likely option once the probabilities are learned. The bandit thus connects the two settings, observing the same subjects while they learn and once the probabilities are known. It also relates to the description–experience gap (Barron and Erev, 2003; Hertwig, Barron, Weber, and Erev, 2004; Hertwig and Erev, 2009; Wulff, Mergenthaler-Canseco, and Hertwig, 2018). The description of the environment is complete and specifies the only channel through which experience should matter, the posterior about the unknown arm. Yet experience affects switching through other channels: over-switching falls when subjects can choose plans that limit their reaction to outcomes.

Implications for the Design of Choice Procedures. The two sources also matter for how choice procedures are designed. Letting people decide in advance when to reconsider, by choosing an option together with how long to stay with it, moves about half of over-switchers to single switching and lengthens their first wave on the unknown option, but leaves most of the others switching. Because the two groups already differ in their free switching while quality is uncertain, who benefits from such a procedure could in principle be anticipated. A deadline, the most common time-based device, targets a different margin: it fixes when the final decision must be made, but does nothing to prevent switching back and forth before it. Because over-switchers interleave draws from the familiar option, they reach a good unknown option's first success later than single-switchers, so a deadline that would suffice for a single-switcher may end their exploration before it pays off. What the results speak to is therefore not the value of a deadline, but that of procedures that limit the reaction to outcomes as they arrive.

Implications for Strategic Experimentation. A question my design does not address, but motivates, is how these individual departures interact with strategic experimentation, in which several agents learn about the same or correlated alternatives. Free-riding leads groups

to under-explore relative to the single-agent benchmark (Keller, Rady, and Cripps, 2005), and Hoelzemann and Klein (2021) provide evidence consistent with this. In their control treatment, where a single agent faces a deterministic familiar arm, most games follow the predicted cutoff rule; with a stochastic familiar arm, about half of the subjects here over-switch and over-explore relative to the same benchmark. When strategic under-exploration and individual over-exploration operate in the same environment, do they offset, or does strategic interaction change the tendency documented here? A first step, which I take in ongoing work, is to vary whether the familiar arm pays a deterministic or a stochastic amount, alone and in a strategic setting.

6 Conclusion

In this paper, I test a central prediction of normative models of sequential experimentation: that switching decisions respond to changes in information about the alternatives' relative qualities. In the one-armed bandit I study, this prediction takes the form of single switching, which restricts choices both before and after the unknown arm (Arm U) pays. I therefore expand the scope of earlier bandit experiments, which study behavior while the quality of at least one alternative is unknown, to behavior once it is known.

I find that switching does not require a change in information, or even uncertainty about the alternatives' qualities. About half of subjects return to an alternative they abandoned without having learned anything new about it, and over-switching persists, though less often, once nothing is left to learn: 39% of subjects switch away from Arm U after it is known to be good. Over-switching is costly: over-switchers explore bad alternatives more, discover good ones later, and earn 20 to 25% less.

In both stages, over-switchers respond to outcomes that carry no information, but only to those in the current wave. I rule out utility curvature as the leading reason for over-switching. Over-switching has two separable sources, present in both stages. When subjects can choose plans that limit their reaction to outcomes, about half of over-switchers adopt single switching, consistent with noisy execution of a correct plan. Most of the rest keep switching, consistent with an incorrect model of the problem or non-standard preferences. Subjects in the two groups differ in their free switching while quality is uncertain but not once it is known; even so, a classification based on the first stage predicts who keeps over-switching in the second.

These results imply that the posterior about the unknown alternative is not enough to

predict switching, suggesting the possibility that returns to discarded alternatives observed in the field, such as rehired employees or reconciled couples, do not always reflect new information. Two questions remain open: what separates an incorrect model of the problem from non-standard preferences among the over-switchers whom choice structure does not correct, and how these tendencies interact with strategic incentives to experiment.

References

Agranov, M., Healy, P. J., & Nielsen, K. (2023). Stable Randomisation. *Economic Journal*, 133(655), 2553–2579.

Anderson, C. M. (2001). *Behavioral Models of Strategies in Multi-Armed Bandit Problems*. Unpublished Ph.D. dissertation, California Institute of Technology.

Anesi, V., & Bowen, T. R. (2021). Policy Experimentation in Committees: A Case against Veto Rights under Redistributive Constraints. *American Economic Journal: Microeconomics*, 13(3), 124–162.

Arnold, J. D., Van Iddekinge, C. H., Campion, M. C., Bauer, T. N., & Campion, M. A. (2021). Welcome Back? Job Performance and Turnover of Boomerang Employees Compared to Internal and External Hires. *Journal of Management*, 47(8), 2198–2225.

Banks, J., Olson, M., & Porter, D. (1997). An Experimental Analysis of the Bandit Problem. *Economic Theory*, 10(1), 55–77.

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

Barlow, M., Doan, T., Friedman, O., & Denison, S. (2025). Probability Matching and Statistical Naïveté. *Judgment and Decision Making*, 20, e37.

Barron, G., & Erev, I. (2003). Small Feedback-Based Decisions and Their Limited Correspondence to Description-Based Decisions. *Journal of Behavioral Decision Making*, 16(3), 215–233.

Benjamin, D. J. (2019). Errors in Probabilistic Reasoning and Judgment Biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics: Applications and Foundations*, Vol. 2 (pp. 69–186). Elsevier.

Bergemann, D., & Välimäki, J. (2008). Bandit Problems. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed., pp. 336–340). Palgrave Macmillan.

Bimpikis, K., Ehsani, S., & Mostagir, M. (2019). Designing Dynamic Contests. *Operations Research*, 67(2), 339–356.

Bland, J. R., & Rosokha, Y. (2024). Rounding the (Non)Bayesian Curve: Unraveling the Effects of Rounding Errors in Belief Updating. Purdue University Economics Working Paper No. 1353.

Bonatti, A., & Hörner, J. (2011). Collaborating. *American Economic Review*, 101(2), 632–663.

Bowen, T. R., Chan, J. M. L., Dube, O., & Lambert, N. S. (2016). Reform Fatigue. Stanford Graduate School of Business Research Paper No. 3394.

Bowen, T. R., Hwang, I., & Krasa, S. (2022). Personal Power Dynamics in Bargaining. *Journal of Economic Theory*, 205, 105530.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Edwards, W. (1961). Probability Learning in 1000 Trials. *Journal of Experimental Psychology*, 62(4), 385–394.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From Anomalies to Forecasts: Toward a Descriptive Model of Decisions under Risk, under Ambiguity, and from Experience. *Psychological Review*, 124(4), 369–409.

Erev, I., & Haruvy, E. (2016). Learning and the Economics of Small Decisions. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics*, Vol. 2 (pp. 638–716). Princeton University Press.

Fan, T. Q., Liang, Y., & Peng, C. (2026). The Inference-Forecast Gap in Belief Updating. *Econometrica*, 94(4), 1279–1312.

Feher da Silva, C., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and Recency as the Main Proximate Causes of Probability Matching: A Reinforcement Learning Analysis. *Scientific Reports*, 7, 15326.

Gaissmaier, W., & Schooler, L. J. (2008). The Smart Potential Behind Probability Matching. *Cognition*, 109(3), 416–422.

Gans, N., Knox, G., & Croson, R. (2007). Simple Models of Discrete Choice and Their Performance in Bandit Experiments. *Manufacturing & Service Operations Management*, 9(4), 383–408.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, 17(3), 295–314.

Gittins, J. C., & Jones, D. M. (1974). A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani, K. Sarkadi, & I. Vincze (Eds.), *Progress in Statistics* (pp. 241–266). North-Holland.

Grohsjean, T., Dokko, G., & Yang, P. (2025). Can You Go Home Again? Performance Assistance Between Boomerangs and Incumbent Employees. *Organization Science*, 36(2), 918–939.

Halac, M., Kartik, N., & Liu, Q. (2016). Optimal Contracts for Experimentation. *Review of Economic Studies*, 83(3), 1040–1091.

Halac, M., Kartik, N., & Liu, Q. (2017). Contests for Experimentation. *Journal of Political Economy*, 125(5), 1523–1569.

Halpern-Meekin, S., Manning, W. D., Giordano, P. C., & Longmore, M. A. (2013). Relationship Churning, Physical Violence, and Verbal Abuse in Young Adult Relationships. *Journal of Marriage and Family*, 75(1), 2–12.

Halpern-Meekin, S., & Turney, K. (2016). Relationship Churning and Parenting Stress Among Mothers and Fathers. *Journal of Marriage and Family*, 78(3), 715–729.

Halpern-Meekin, S., & Turney, K. (2023). Romantic Unions and Mental Health: The Role of Relationship Churning. *Journal of Health and Social Behavior*, 64(2), 243–260.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8), 534–539.

Hertwig, R., & Erev, I. (2009). The Description–Experience Gap in Risky Choice. *Trends in Cognitive Sciences*, 13(12), 517–523.

Hoelzemann, J., & Klein, N. (2021). Bandits in the Lab. *Quantitative Economics*, 12(3), 1021–1051.

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92(5), 1644–1655.

Hörner, J., Klein, N., & Rady, S. (2022). Overcoming Free-Riding in Bandit Games. *Review of Economic Studies*, 89(4), 1948–1992.

Horowitz, A. V. (1973). *Experimental Study of the Two-Armed Bandit Problem*. Unpublished Ph.D. dissertation, University of North Carolina, Chapel Hill.

Hudja, S., & Woods, D. (2024a). Exploration versus Exploitation: A Laboratory Test of the Single-Agent Exponential Bandit Model. *Economic Inquiry*, 62(1), 267–286.

Hudja, S., & Woods, D. (2024b). A Short Survey on Bandit Experiments. Working paper.

Hudja, S., & Woods, D. (2025). Strategies in the Multi-Armed Bandit. *Experimental Economics*, 28(4), 974–996.

Keller, G., Rady, S., & Cripps, M. (2005). Strategic Experimentation with Exponential Bandits. *Econometrica*, 73(1), 39–68.

Keller, G., & Rady, S. (2010). Strategic Experimentation with Poisson Bandits. *Theoretical Economics*, 5(2), 275–311.

Khromenkova, D. (2015). Collective Experimentation with Breakdowns and Breakthroughs. Working paper, University of Mannheim.

Kim, D. G., & Kim, H. C. (2022). Probability Matching and Strategic Decision Making. *Journal of Behavioral and Experimental Economics*, 98, 101850.

Koehler, D. J., & James, G. (2010). Probability Matching and Strategy Availability. *Memory & Cognition*, 38(6), 667–676.

Kreps, D. M. (1979). A Representation Theorem for "Preference for Flexibility". *Econometrica*, 47(3), 565–578.

Kumar, V., Bhagwat, Y., & Zhang, X. (2015). Regaining "Lost" Customers: The Predictive Power of First-Lifetime Behavior, the Reason for Defection, and the Nature of the Win-Back Offer. *Journal of Marketing*, 79(4), 34–55.

Kumar, V., Leszkiewicz, A., & Herbst, A. (2018). Are You Back for Good or Still Shopping Around? Investigating Customers' Repeat Churn Behavior. *Journal of Marketing Research*, 55(2), 208–225.

McMahon, A. J., & Scheel, M. H. (2010). Glucose Promotes Controlled Processing: Matching, Maximizing, and Root Beer. *Judgment and Decision Making*, 5(6), 450–457.

Meyer, R. J., & Shi, Y. (1995). Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. *Management Science*, 41(5), 817–834.

Monk, J. K., Ogolsky, B. G., & Oswald, R. F. (2018). Coming Out and Getting Back In: Relationship Cycling and Distress in Same- and Different-Sex Relationships. *Family Relations*, 67(4), 523–538.

Newell, B. R., Koehler, D. J., James, G., Rakow, T., & van Ravenzwaaij, D. (2013). Probability Matching in Risky Choice: The Interplay of Feedback and Strategy Availability. *Memory & Cognition*, 41(3), 329–338.

Newell, B. R., & Rakow, T. (2007). The Role of Experience in Decisions from Description. *Psychonomic Bulletin & Review*, 14(6), 1133–1139.

Rabin, M. (2002). Inference by Believers in the Law of Small Numbers. *Quarterly Journal*

of Economics, 117(3), 775–816.

Rabin, M., & Vayanos, D. (2010). The Gambler's and Hot-Hand Fallacies: Theory and Applications. *Review of Economic Studies*, 77(2), 730–778.

Robbins, H. (1952). Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, 58(5), 527–535.

Roth, A. E., & Murnighan, J. K. (1978). Equilibrium Behavior and Repeated Play of the Prisoner's Dilemma. *Journal of Mathematical Psychology*, 17(2), 189–198.

Saha, A. (1993). Expo-Power Utility: A "Flexible" Form for Absolute and Relative Risk Aversion. *American Journal of Agricultural Economics*, 75(4), 905–913.

Scharfstein, D. O., Tsiatis, A. A., & Robins, J. M. (1997). Semiparametric Efficiency and Its Implication on the Design and Analysis of Group-Sequential Studies. *Journal of the American Statistical Association*, 92(440), 1342–1350.

Schulze, C., James, G., Koehler, D. J., & Newell, B. R. (2019). Probability Matching Does Not Decrease under Cognitive Load: A Preregistered Failure to Replicate. *Memory & Cognition*, 47(3), 511–518.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A Re-examination of Probability Matching and Rational Choice. *Journal of Behavioral Decision Making*, 15(3), 233–250.

Strulovici, B. (2010). Learning While Voting: Determinants of Collective Experimentation. *Econometrica*, 78(3), 933–971.

Swider, B. W., Liu, J. T., Harris, T. B., & Gardner, R. G. (2017). Employees on the Rebound: Extending the Careers Literature to Include Boomerang Employment. *Journal of Applied Psychology*, 102(6), 890–909.

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4), 285–294.

Vulkan, N. (2000). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, 14(1), 101–118.

West, R. F., & Stanovich, K. E. (2003). Is Probability Matching Smart? Associations between Probabilistic Choices and Cognitive Ability. *Memory & Cognition*, 31(2), 243–251.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A Meta-Analytic Review of Two Modes of Learning and the Description-Experience Gap. *Psychological Bulletin*,

144(2), 140–176.

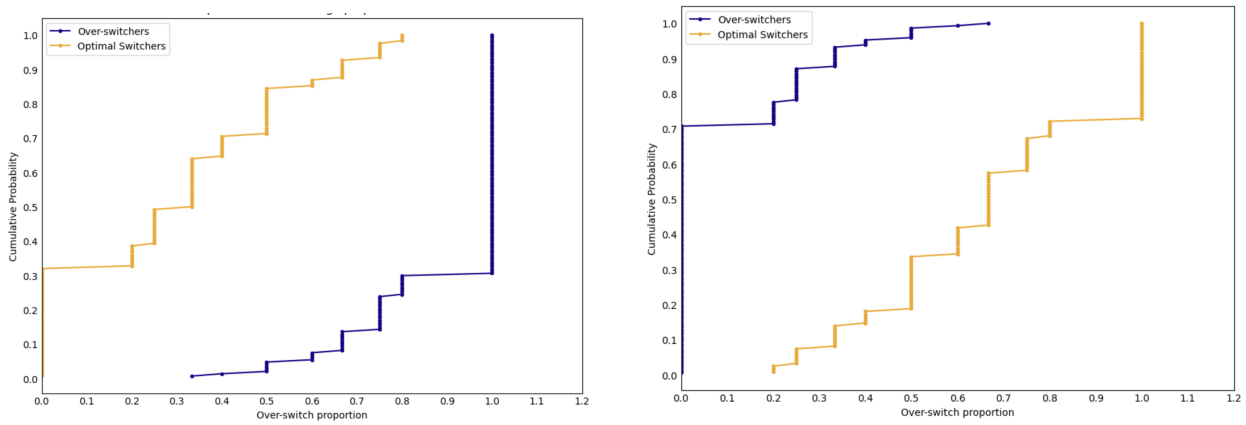
7 Appendix

7.1 Figures and Tables

Table 3: Switching behavior across segments under alternative definitions

	(0) Main	(1) All	(2) Problems	(3)	(4)	(5)
				Problem 5, whole problem		
	Text	problems	4–5	All	A paid	A never paid
Fully optimal	98 (37.7)	91 (35.0)	49 (36.8)	143 (47.7)	68 (56.7)	75 (41.7)
Over in some segments only	69 (26.5)	111 (42.7)	58 (43.6)	29 (9.7)	29 (24.2)	—
Over in all segments	83 (31.9)	48 (18.5)	22 (16.5)	106 (35.3)	23 (19.2)	83 (46.1)
Under-switching present	10 (3.8)	10 (3.8)	4 (3.0)	22 (7.3)	0 (0.0)	22 (12.2)
Segments classified	2	3	3	1–2	2	1
Observations	260	260	133	300	120	180

Notes: Cells report the number of subjects and, in parentheses, the percentage of the column total. Segments are the QU stage (before Arm U first pays) and the QC stage (after Arm U first pays). Each segment is classified as optimal, over-switching or under-switching relative to the single-switch cutoff rule; in the QC stage, any switch away from Arm U counts as over-switching. Column (0) uses the definition in the main text: it classifies two segments per subject, the QU stage of the last problem in which Arm U never paid and the QC stage of the last problem in which Arm U paid, with the same sample restrictions as column (1). Columns (1) and (2) classify three segments per subject: the QU stage of the last problem in which Arm U never paid, and the QU and QC stages of the last problem in which Arm U paid; column (1) excludes subjects with no problem in which Arm U never paid (5), no problem in which Arm U paid (33), or a one-round QC stage (2); column (2) further restricts to subjects for whom both classified problems are Problem 4 or 5, i.e. whose outcomes in Problems 4 and 5 differ. Columns (3)–(5) classify Problem 5 as a whole: subjects paid by Arm U have a QU and a QC segment; subjects never paid by Arm U have a single segment covering the whole problem. A segment that is undefined for a subject is not used in that subject's classification. In particular, the 16 subjects in columns (3) and (4) for whom Arm U paid in the first round have no QU segment and are classified on their QC stage alone. "Fully optimal": every classified segment is optimal. "Over in some segments": at least one segment is over-switching and at least one is optimal, with no under-switching. "Over in all segments": every classified segment is over-switching. "Under-switching present": at least one segment is under-switching. Under-switching cannot occur in the QU stage of a problem where Arm U paid, since it requires never trying Arm U.



(a) Proportion of problems with over-switching

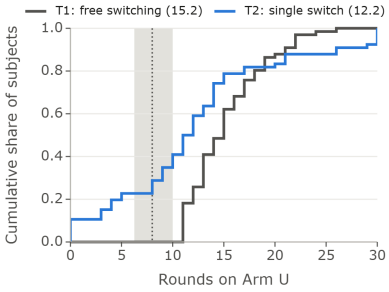
(b) Proportion of problems with optimal-switching

Figure 5: Cumulative Distribution Function of Problems across Treatment 1. Only problems with at least 5 rounds in the QU informational stage were included.

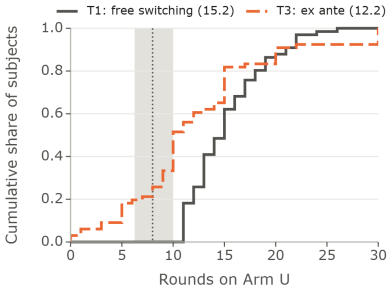
Table 4: Kolmogorov–Smirnov Tests of Switching Behavior

	KS Statistic	p-value
Over-switch proportion	0.7966	< 0.001
Optimal-switch proportion	0.7970	< 0.001

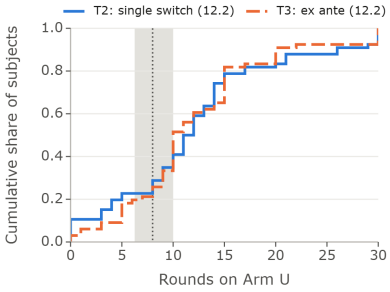
Note: Only problems with at least 5 rounds in the *QU* informational stage were included.



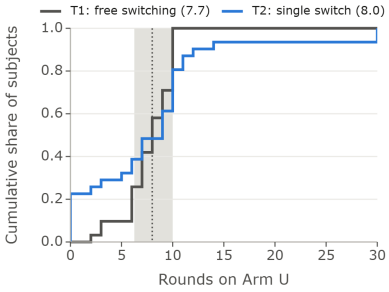
(a) Treatment 1 versus Treatment 2



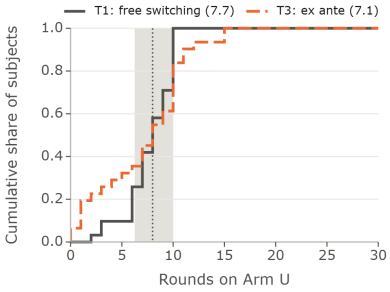
(b) Treatment 1 versus Treatment 3



(c) Treatment 2 versus Treatment 3



(d) Treatment 1 versus Treatment 2



(e) Treatment 1 versus Treatment 3

Figure 6: Exploration of Treatment 1 Over-Switchers. The top row shows over-switchers who explored more than the single-switchers' 75th percentile in Treatment 1 (more than 10 choices of Arm U, $n = 66$), while the bottom row shows over-switchers at or below it ($n = 31$). The shaded area is the single-switchers' interquartile range in Treatment 1 (6.25 to 10 rounds); the dotted line marks the risk-neutral benchmark of eight rounds. Means in parentheses.

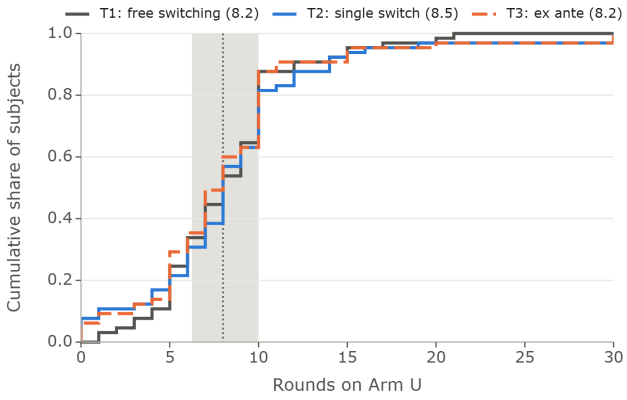


Figure 7: Exploration of Treatment 1 single-switchers across treatments ($n = 65$). The shaded area is the single-switchers' interquartile range in Treatment 1 (6.25 to 10 rounds); the dotted line marks the risk-neutral benchmark of eight rounds. Means in parentheses.

Table 5: Kolmogorov–Smirnov Tests of Exploration Across Treatments

	KS Statistic	p-value
Treatment 1 over-switchers exploring more than 10 rounds ($n = 66$)		
Treatment 1 vs. Treatment 2	0.4091	< 0.001
Treatment 1 vs. Treatment 3	0.5152	< 0.001
Treatment 2 vs. Treatment 3	0.1061	0.8556
Treatment 1 over-switchers exploring 10 rounds or fewer ($n = 31$)		
Treatment 1 vs. Treatment 2	0.2258	0.4135
Treatment 1 vs. Treatment 3	0.2258	0.4135
Treatment 2 vs. Treatment 3	0.1613	0.8235
Treatment 1 single-switchers ($n = 65$)		
Treatment 1 vs. Treatment 2	0.0769	0.9916
Treatment 1 vs. Treatment 3	0.0615	0.9998
Treatment 2 vs. Treatment 3	0.1077	0.8493

Table 6: Exploration under Treatments 2 and 3 relative to each subject's Treatment 1 average

			Treatment 2			Treatment 3		
Treatment 1 group	n	T1 average	Mean	Change (p)	% reducing	Mean	Change (p)	% reducing
Single-switchers	65	8.54	8.52	−0.02 (0.971)	45	8.15	−0.39 (0.483)	57
Over-switchers, more than 10 rounds	66	14.92	12.23	−2.69 (0.009)	67	12.18	−2.74 (0.005)	71
Over-switchers, 10 rounds or fewer	31	9.02	7.97	−1.05 (0.46)	52	7.13	−1.89 (0.032)	58

Notes: Sessions 1 and 2. Groups are defined in the last Treatment 1 problem in which Arm U never paid; 10 rounds is the single-switchers' 75th percentile of exploration in Treatment 1. T1 average: each subject's mean number of Arm U choices over all their Treatment 1 problems in which Arm U never paid, including the one used to define the groups. Treatment 2: last Treatment 2 problem in which Arm U never paid; Treatment 3: the executed ex ante report. Change: mean of each subject's difference from their T1 average, with the p-value of a paired t-test. % reducing: share of subjects below their T1 average.

	QU Single-Switchers			Single-Switchers in Both Stages		
	Exploration T1			Exploration T1		
	Below IQR	Within IQR	Above IQR	Below IQR	Within IQR	Above IQR
Same part in T2 and T3	14	22	4	8	19	3
Moves in T2 or T3	8	13	4	7	12	4

Table 7: Changes in exploration for single-switchers. IQR: the single-switchers' interquartile range of exploration in Treatment 1 (6.25 to 10 rounds).
Same part: below, within or above the IQR in both T2 and T3, as in Treatment 1; moves: in a different part in at least one of them.

7.2 Robustness: Two Checks Using Session 2

Session 2 differs from Session 1 in two respects that provide useful robustness checks. First, it included only 2 Treatment 2 problems instead of the 5 used in Session 1. Second, it included an additional free-switching problem after Treatment 3.

This additional free-switching problem allows me to examine whether order or learning effects account for the treatment results. In it, 52% of problems in which Arm U never paid deviate from the single-switching benchmark, the same share as in Treatment 1 for these subjects. Among the 54 subjects who explored Arm U in both problems, 41 either single-switch or over-switch in both treatments, with the remaining subjects roughly balanced between improving and worsening. More importantly, the difference in exploration between single- and over-switchers persists: 8.8 versus 13.4 rounds ($p < 0.01$). Thus, the treatment effects are unlikely to be driven primarily by order or learning among Session 2 subjects.

Among subjects who over-switched and explored beyond the single-switchers' 75th percentile in Treatment 1, exploration in Treatments 1 and 3 is similar across sessions, whereas Treatment 2 differs markedly: 15.0 rounds in Session 1 versus 8.3 in Session 2. This difference may reflect the shorter exposure to Treatment 2 in Session 2 (2 problems versus 5 in Session 1).[41]

7.3 Treatment 5: Decision Rules Details

The rules distinguish between two phases. Phase 1 goes from the start of the problem until Arm U returns a reward, if that ever happens. Phase 2 goes from the moment Arm U returns a reward until the end of the problem. Rules 2, 3, and 4 apply to Phase 1 only. Separately, subjects choose whether, in the case of reaching Phase 2, they would rather select Arm U for the rest of the problem or keep their Phase 1 rule, with the option of updating the parameters.

Rule 1: *Single Switching.* The participant chose a single number of rounds for which the computer would initially draw from Jar U. If Jar U produced no black ball during this period, the computer permanently switched to Jar F for the remainder of Phase 1. If a black ball was observed before the scheduled switch, the computer remained on Jar U until the end of the problem.

Rule 2: *Consecutive White-Ball Draws.* The participant chose separate thresholds for Jar U and Jar F. The computer switched away from the current jar once the corresponding number of consecutive white-ball draws had been observed. The computer therefore alternated between the two jars whenever the relevant threshold was reached, with the thresholds fixed throughout Phase 1.

Rule 3: *Consecutive White-Ball Draws with Adaptive Thresholds.* This rule followed the same outcome-based switching mechanism as Rule 2, but allowed the participant to choose a new threshold each time the computer returned to a jar. Thus, switching thresholds could vary across successive waves on the same jar.

Rule 4: *Cycles.* The participant chose the number of rounds for which the computer would remain on each jar. The computer then alternated between Jar U and Jar F after

[41] The two subject pools may differ. However, the similarity of the other treatments across sessions provides some evidence against a general session-level difference, though it does not rule out differences specific to Treatment 2.

every fixed number of rounds, regardless of the outcomes observed. The cycle length was fixed throughout Phase 1.

Thus, the rules vary along two dimensions: whether switching is based on observed outcomes or elapsed rounds, and whether the switching threshold is fixed or can be adjusted across waves on a jar.

The rules were presented in two randomized orders: (Rule 4, Rule 1, Rule 2, and Rule 3) and (Rule 2, Rule 3, Rule 1, and Rule 4).

7.4 Empirical Analyses

Welfare Analysis

The main-text welfare comparison excludes the 26 subjects who never explored Arm U. They earn the most in problems in which Arm U never paid (157.69 points), but only because Arm U happened not to pay: a subject who never explores can never benefit from Arm U when it is good.

The payoff distributions of single-switchers and over-switchers also differ (Kolmogorov–Smirnov $p = 0.030$): the median single-switcher earns 100 points, against 50 for over-switchers.

Table 8: Payoffs and Switching, Problems in Which Arm U Never Paid: OLS regression

	Full Payoff
Switch number	-2.550^{***}
	(0.632)
Started with Arm U	4.914
	(10.483)
Constant	100.533^{***}
	(10.298)
Observations	269
R^2	0.038

Robust (HC1) standard errors in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Welfare Analysis in the last problem of Treatment 1 Considering only Problem 5 of Treatment 1, among subjects for whom Arm U never paid, single-switchers earn 112.00

Table 9: Payoffs and Switching, Problems in Which Arm U Paid: OLS regression

	(1) Full Payoff	(2) Full Payoff
Switch number (QU stage)	-11.031^{***}	-0.667
	(2.012)	(2.279)
Switch number (QC stage)	-2.518	-4.253^{**}
	(1.914)	(1.932)
Rounds remaining after first reward		7.606^{***}
		(1.244)
Constant	299.979^{***}	116.353^{***}
	(8.814)	(28.359)
Observations	228	228
R^2	0.074	0.162

Robust (HC1) standard errors in parentheses. Rounds remaining after first reward equal 29 minus the round of the first reward from Arm U. QU-stage switches and rounds remaining after the first reward are correlated at -0.53; each QU-stage switch delays the first reward by 1.36 rounds ($p < 0.001$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

points and over-switchers 77.11 points ($p = 0.003$; Holm-adjusted $p = 0.006$). In a regression of Problem 5 payoffs on QU-stage over-switching, an indicator for whether Arm U paid, their interaction and a QC-stage deviation indicator (Table 12, column 2), QU-stage over-switching is associated with lower payoffs both when Arm U never paid (-34.89 points, $p = 0.003$) and when it paid (-44.10 points, $p = 0.096$); the two effects are not significantly different ($p = 0.75$). The QC-stage deviation estimate has the expected sign but is imprecise with only 119 subjects reaching that stage (-31.61 points, $p = 0.18$). Pooling both states, subjects classified as single-switchers over the whole of Problem 5 earn 204.20 points against 142.96 for over-switchers ($p < 0.001$). Over-switchers are also 32.7 percentage points less likely to have been paid by Arm U (column 3), so column (2) measures the cost of over-switching conditional on whether Arm U paid.

7.5 Robustness to selection into problems in which Arm U paid or never paid

Whether Arm U paid at some point in a problem or never did depends partly on subjects' own behavior: a subject who stops sampling Arm U may never observe a reward from a good arm, and a subject who samples it persistently is more likely to find one. Comparisons

Table 10: Pace of sampling Arm U before the first reward

	(1) OLS	(2) Median regression
Draws from Arm U until first reward	1.023***	1.000
	(0.063)	(0.000)
QU-stage over-switcher	1.784*	−0.333
	(1.074)	(1.007)
Draws × QU-stage over-switcher	0.475**	0.778***
	(0.188)	(0.228)
Constant	0.824***	0.000
	(0.312)	(0.000)
Observations	223	223
R^2 / pseudo-R^2	0.676	0.625

Notes: Dependent variable: round of the problem in which Arm U first paid. Sample: QU-stage single-switchers and over-switchers in their last problem in which Arm U paid, if that problem had at least one unsuccessful draw from Arm U, the minimum required to over-switch. Draws from Arm U count the rounds on Arm U up to and including the first reward. Column (1): OLS with heteroskedasticity-robust (HC1) standard errors. Column (2): median regression with standard errors from 1,000 bootstrap resamples of subjects; the intercept and the slope on draws are estimated without sampling variation because most single-switchers lie exactly on the line where the round of the first reward equals the number of draws. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

between problems in which Arm U paid and those in which it never paid could therefore reflect selection into problem type rather than behavior. To address this, I use the quality of Arm U assigned by the experiment in each problem (bad arms never pay; good arms pay with probability 0.2), which is independent of subjects' choices. Quality vectors are available for all 300 subjects. In all 1,500 subject-problems, a bad Arm U never paid, with a single exception attributable to the reconstruction of rewards in rounds 29 and 30, in a problem other than Problem 5.

I focus on Problem 5, which all subjects face, and classify subjects in two ways: by their behavior before the first reward (QU stage), and by their behavior over the whole problem (QU and QC stages together; a subject is classified as an over-switcher if they over-switch in at least one stage). When Arm U is bad there is no QC stage and both classifications coincide.

Main results. Table 13 reports the results. Over-switching is prevalent under both qualities: 49.7% of subjects over-switch when Arm U is bad and 39.9% when it is good. When Arm U is bad, over-switchers spend 13.36 rounds on it against 9.13 for single-switchers

Table 11: Probability that a draw from Arm U pays, last problem in which Arm U paid

	Not yet over-switched		Already over-switched		
Draw from Arm U	Pay rate	Draws	Pay rate	Draws	p-value
1st	0.165	267	—	—	—
2nd	0.238	185	0.289	38	0.537
3rd	0.226	133	0.257	35	0.660
4th	0.216	97	0.156	32	0.613
5th	0.306	72	0.419	31	0.268
6th	0.245	49	0.421	19	0.235
7th	0.333	36	0.333	12	1.000
8th	0.042	24	0.375	8	0.039
Difference (already – not yet), draws 2–15: 0.051 (0.033), $p = 0.122$					
Draws: 1,143 (267 subjects); draws 2–15 used in the difference: 864 (223 subjects)					

Notes: Pay rate is the share of QU-stage draws from Arm U that returned a reward, among subjects making that draw (i.e. conditional on all earlier draws having failed). A subject has already over-switched at a given draw if, before it, they had left Arm U and returned to it. Row p-values are from Fisher's exact test of equal pay rates at that draw; with seven row-level tests, the single $p < 0.05$ (8th draw, based on 8 draws) does not survive a correction for multiple comparisons. The difference is the coefficient on the over-switching indicator in a linear regression of whether the draw paid on draw-number dummies, with standard errors clustered by subject; the first draw is excluded because no subject can have over-switched before it.

($p < 0.001$) and earn 72.44 points against 106.56 ($p = 0.004$). When Arm U is good, single-switchers and over-switchers are about equally likely to find its reward (84.9% and 83.3%), but over-switchers find it later (round 11.64 against 5.62, $p = 0.001$) and earn less: 276.83 against 239.47 points using the whole-problem classification (Welch $p = 0.071$; Mann–Whitney $p = 0.053$), and 267.74 against 218.33 using the QU-stage classification ($p = 0.038$). Since the quality of Arm U is randomly assigned, none of these patterns can be driven by selection into problems in which Arm U paid or never paid.

Timing of the first reward and classification. When Arm U is good, the QU-stage classification is itself related to when the first reward arrives: a subject paid in the first two or three rounds has no opportunity to leave Arm U and return to it, and is therefore classified as a QU-stage single-switcher. Table 14 shows that the share of QU-stage over-switchers rises from 0% among subjects paid in the first round to 52% among those paid after round 10. Early finders, however, have more rounds in which to deviate after the first reward, and they do so at a similar rate as other subjects: QC-stage deviation rates lie between 30% and 45% regardless of when the reward arrived. As a result, the whole-problem classification

Table 12: Payoffs and Switching in Problem 5: OLS regressions

	(1) Full Payoff	(2) Full Payoff	(3) Arm U paid
QU-stage over-switcher	-100.566^{***}	-34.892^{***}	-0.327^{***}
	(14.331)	(11.638)	(0.056)
QU-stage under-switcher	-58.984^{***}	42.545^{**}	-0.559^{***}
	(20.973)	(20.197)	(0.038)
Arm U paid		190.668^{***}	
		(15.979)	
QU-stage over-switcher × Arm U paid		-9.203	
		(28.917)	
QC-stage deviation		-31.609	
		(23.319)	
Constant	213.529^{***}	112.000^{***}	0.559^{***}
	(10.340)	(8.457)	(0.038)
Observations	300	299	300
R^2	0.134	0.509	0.151

Problem 5, all subjects. Robust (HC1) standard errors in parentheses. The omitted category is QU-stage single-switchers. Arm U paid equals one if Arm U returned a reward at some point in the problem. QC-stage deviation equals one for subjects paid by Arm U who switched away from it at least once after the first reward, and zero otherwise. One subject paid by Arm U in the last round has no QC stage and is excluded from column (2). Subjects for whom Arm U paid in the first round (16) have no QU stage and are coded as not deviating in it. Column (3) is a linear probability model. No under-switcher was paid by Arm U, so the under-switcher coefficient in column (2) is estimated only among subjects for whom Arm U never paid. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

is largely unrelated to timing (33–44% over-switchers among subjects paid by round 10). When Arm U is bad no subject is ever paid, so the classification reflects behavior alone.

Last problems in which Arm U paid or never paid. All 267 last problems in which Arm U paid had a good Arm U, since a bad arm never pays. Among the 295 last problems in which Arm U never paid, 34 (11.5%) had a good Arm U that the subject never found. Arm U was bad in 86.1% of single-switchers' last problems in which it never paid and in 92.5% of over-switchers' (Fisher exact $p = 0.108$). Restricting the comparison to problems in which Arm U was bad, single-switchers earn 102.38 points and over-switchers 75.00 ($p = 0.002$), close to the gap among all last problems in which Arm U never paid (107.38 against 80.61).

Table 13: Problem 5 by the quality of Arm U

	Single	Over	Under	Difference	Welch p	M–W p
Panel A. Arm U bad (157 subjects)						
Subjects (% of panel)	61 (38.9)	78 (49.7)	18 (11.5)			
Rounds on Arm U	9.13	13.36	0.00	−4.23	< 0.001	< 0.001
Share of rounds on Arm U	0.30	0.45	0.00			
Switches	1.08	7.95	0.00			
Payoff (points)	106.56	72.44	152.78	34.12	0.004	0.001
Panel B. Arm U good (143 subjects), classification over QU and QC						
Subjects (% of panel)	82 (57.3)	57 (39.9)	4 (2.8)			
Payoff (points)	276.83	239.47	162.50	37.36	0.071	0.053
Panel C. Arm U good, QU-stage classification						
Subjects	93	30	4			
Found the reward (%)	84.9	83.3	0.0			
Round of first reward (finders)	5.62	11.64	—	−6.02	0.001	< 0.001
Draws from Arm U needed (finders)	4.59	6.48	—	−1.89	0.093	0.075
Payoff (points)	267.74	218.33	162.50	49.41	0.038	0.037

Notes: Problem 5, all 300 subjects. Single, Over and Under refer to single-, over- and under-switchers. Differences are single-switchers minus over-switchers, tested with Welch t-tests and Mann–Whitney (M–W) tests. In Panel C, 16 subjects paid by Arm U in the first round have no QU stage and are not classified; in Panel B they are classified on their QC-stage behavior. Rounds on Arm U are counted over the whole problem when Arm U is bad. Draws needed among finders partly reflect classification: a subject paid within the first few draws cannot become a QU-stage over-switcher (see text).

7.6 Responses to Outcomes: Data, Inference and Robustness

This appendix accompanies Table 1. Unless stated otherwise, the sample is the 147 QU-stage over-switchers (Arm F and Arm U in the QU stage) or the 103 QC-stage over-switchers (the

Table 14: Over-switching and the timing of the first reward, Problem 5, Arm U good

First reward in	Subjects	Over-switcher in QU	Deviated in QC	Over-switcher overall
Round 1	16	0.0%	37.5%	37.5%
Rounds 2–3	32	9.4%	37.5%	43.8%
Rounds 4–6	33	15.2%	30.3%	33.3%
Rounds 7–10	18	33.3%	33.3%	38.9%
Round 11 or later	21	52.4%	45.0%	66.7%

Notes: Subjects who found the reward from a good Arm U in Problem 5 and are classified as single- or over-switchers over the whole problem. "Over-switcher overall" combines the QU and QC stages. QC-stage deviation shares are computed among subjects with a classifiable QC stage.

QC stage), in all five Treatment 1 problems, and estimates are linear probability models in percentage points with standard errors clustered by subject.

Data and definitions

Decisions. A decision is the choice in round $t+1$ after the outcome of round t, for $t = 1, \ldots, 28$; the position in the wave is the number of consecutive rounds on the current arm up to and including round t. A decision belongs to the QU stage if Arm U has not paid in rounds $1, \ldots, t$, and to the QC stage otherwise; before Arm U pays, the number of failures on Arm U equals the number of rounds on it.

Rounds 29 and 30. The experiment's code does not separate the outcomes of rounds 29 and 30 (footnote 19): when the subject chooses the same arm in both rounds, a reward in round 30 alone is never stored, and on Arm F a round-29 reward is stored in 20.8% of cases, close to the probability of at least one reward in two rounds ($1 - 0.9^2 = 19\%$). When the subject switches arms between the two rounds, one reward is stored in 146 problems, against about 27 expected.[42]

Draws. Outcomes match the design probabilities on which the randomization tests rely: Arm F pays in 9.9% of rounds and a good Arm U in 19.7%, independently of the previous outcome, the previous arm and, for Arm F, the round ($p \geq 0.29$); the rate of a good Arm U varies somewhat across rounds ($p = 0.035$), but round fixed effects leave rows 5 and 6 of Table 1 unchanged (-5.93, $p = 0.008$; -3.13, $p < 0.001$).

[42]Because a round-29 payment from Arm U may not be stored when the subject switches to Arm F in round 30, three problems recorded as ones in which Arm U never paid could have had an Arm U payment in round 29. None of them is the problem used to classify a subject in the QU stage; two belong to subjects whose last problem in which Arm U paid comes later, and the third to a subject with no problem in which Arm U paid, who is therefore not classified in the QC stage.

Samples. The 148 classified subjects who are not QU-stage over-switchers are 122 single-switchers and 26 subjects who never tried Arm U in their last problem in which Arm U never paid; five subjects have no problem in which Arm U never paid and are not classified. Problems with Arm U bad contain 76% of the QU-stage Arm F decisions of over-switchers, so the Arm U bad estimates are a large subsample of the main sample rather than an independent replication.

Estimation and inference

Linear models. All fixed effects enter as indicators. Standard errors are clustered by subject with the usual small-sample correction, and joint tests are Wald tests with the clustered covariance matrix. Groups of subjects are compared with separate regressions, each with its own coefficients and fixed effects, so the standard error of a difference is $\sqrt{se_1^2 + se_2^2}$.

Logit models. The fixed-effects logit uses the same fixed effects as the linear model and estimates them as parameters. Decisions in fixed-effect cells in which the outcome never varies have a fitted probability of zero or one and contribute a marginal effect of zero. Average marginal effects are averaged over all decisions of the sample, so that they are comparable with the linear estimates. QU-stage over-switchers make on average 49 Arm F decisions, which limits the incidental-parameter bias of the subject effects.

Randomization tests. Arm F and Arm U draws are i.i.d., so under the sharp null hypothesis that choices do not depend on outcomes, every choice would have been the same under any other realization of the draws. I redraw every outcome of the arm and stage under study with its design probability, hold all choices fixed, recompute every regressor and every control that depends on outcomes, and re-estimate the regression. The randomization p-value is the share of redraws (5,000 for Arm F, 3,000 for Arm U in the QC stage) whose coefficient is at least as large in absolute value as the observed one. The test is exact in finite samples and needs no assumption about how subjects select into positions in the wave: choices, and therefore selection, are the same in every redraw. Row 3 has no randomization test, because the null hypothesis it tests (curvature of utility over earnings) allows rewards to affect choices.

The problem used for classification. The analyses include the five Treatment 1 problems and are conditional on subjects being classified as over-switchers in their last problem in which Arm U never paid for the QU stage and in their last problem in which Arm U paid for the QC stage, as in the main analysis.

Logit estimates and randomization tests

Table 15 repeats the tests of Table 1 with other estimators (Panel A) and on other samples (Panel B). Panel A reports, for every row of Table 1 and for switching back to Arm U before the end of the problem (Table 17), the logit average marginal effect, the randomization p-value and the estimate when Arm U is bad. The logit checks that the linear model, with many fixed effects and a low rate of switching, does not drive the estimates; the randomization test checks inference without relying on the estimates being causal or on large-sample approximations. The two models give the same sign in every row, and their p-values lie on the same side of 0.10 in every row.

Alternative definitions of over-switchers

The first three columns of Panel B of Table 15 repeat the main tests under the alternative definitions of over-switchers used in the model comparison, because the main definition rests only on the last problem in which Arm U never paid (QU stage) or the last problem in which Arm U paid (QC stage). The column "last U paid or never paid" uses the problems of the main definition but pools the two stages: a subject over-switches if he does so in either (175 subjects, including the 147). Under "last two problems", over-switchers are a subset of those, counting these problems only when they are Problem 4 or 5 (147 subjects, 133 of whom are over-switchers under the main definition). Under "last problem", subjects are classified on their behavior in the very last problem, Problem 5, in either stage (135 subjects, 123 of whom are over-switchers under the main definition); unlike the "Problem 5 alone" column, the estimates use all five problems. The main results hold under every definition, with two exceptions: the effect of a single Arm U reward in the QC stage (row 5) has $p = 0.050$ under "last problem", and when decisions are also matched on round, total earnings and failures on Arm U (row 3), the estimates keep their sign but are significant at the 5% level only without subject fixed effects under "last U paid or never paid" ($p = 0.048$).

Full population and Problem 5

The "Full population" and "Problem 5 alone" columns of Table 15 (Panel B) repeat the main tests on all subjects and on Problem 5 alone. The full population checks that the results do not depend on how over-switchers are classified; Problem 5 checks them within a single problem, with QU-stage over-switchers defined by their QU-stage behavior in that problem and QC-stage over-switchers by their QC-stage behavior in it, as in the main analysis. In the

Table 15: Robustness of Table 1: estimators and samples

Panel A. Estimators	Linear	Logit	Rand. p	Arm U bad
Reward → switches back to Arm U	−4.58	−3.21	<0.001	−5.35
	[0.002]	[0.013]		[0.002]
(1) Reward in the round just played	−0.40	−0.38	–	−0.07
	[0.849]	[0.854]		[0.974]
(2) Rewards in current wave	−3.10	−3.81	<0.001	−3.40
	[0.002]	[0.002]		[0.002]
(3a) State held fixed	−2.79	−3.74	–	−2.52
	[0.050]	[0.027]		[0.185]
(3b) State held fixed, subject	−3.47	−4.29	–	−3.99
	[0.034]	[0.013]		[0.047]
(4) Current wave	−3.15	−3.71	<0.001	−3.35
	[0.004]	[0.003]		[0.003]
Earlier waves	+0.74	+0.72		+0.74
	[0.528]	[0.531]		[0.595]
Difference	−3.88	−4.44		−4.09
	[0.013]	[0.006]		[0.023]
(5) Arm U reward just played (QC)	−6.02	−6.21	–	
	[0.007]	[0.009]		
(6) Current-wave Arm U rewards (QC)	−3.24	−4.41	<0.001	
	[<0.001]	[<0.001]		

Panel B. Samples	Definition of over-switchers			Subjects, problems or waves		
	Last U paid or never paid	Last two problems	Last problem	Full population	Problem 5 alone	Returning waves
Reward → switches back to Arm U	−5.54	−4.92	−4.80	−3.70	−4.38	–
	[<0.001]	[0.001]	[0.001]	[<0.001]	[0.112]	
(2) Rewards in current wave	−2.92	−3.36	−3.42	−1.27	−4.19	−2.65
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[0.025]	[0.055]
(3a) State held fixed	−2.47	−2.14	−2.40	−1.61	−2.94	−2.43
	[0.048]	[0.107]	[0.090]	[0.062]	[0.533]	[0.242]
(3b) State held fixed, subject	−2.76	−2.60	−3.08	−2.08	−6.81	−3.68
	[0.063]	[0.111]	[0.078]	[0.043]	[0.422]	[0.114]
(4) Current minus earlier waves	−3.48	−3.59	−3.78	−2.09	−5.40	−3.69
	[0.010]	[0.014]	[0.014]	[0.027]	[0.278]	[0.039]
(5) Arm U reward just played (QC)	−2.83	−3.13	−3.44	−1.78	−8.79	–
	[0.022]	[0.042]	[0.050]	[0.015]	[0.019]	
(6) Current-wave Arm U rewards (QC)	−1.43	−1.66	−2.00	−0.68	−9.03	–
	[<0.001]	[0.001]	[0.001]	[<0.001]	[<0.001]	
Subjects (row 2)	174	146	134	299	99	147

Notes: Coefficients in percentage points, p-values in brackets, standard errors clustered by subject. Rows, samples and controls as in Table 1; switching back to Arm U before the end of the problem as in Table 17. Panel A. Linear: linear probability model, the estimates of Table 1 and Table 17. Logit: fixed-effects logit average marginal effect over all decisions. Rand. p: randomization p-value, two-sided, from 5,000 redraws of every QU-stage Arm F outcome or, for row 6, 3,000 redraws of every QC-stage Arm U outcome, choices held fixed; "<0.001" means at most 3 of 5,000 or none of 3,000 redraws as extreme as the observed coefficient; rows marked – were not tested. Arm U bad: linear probability model on the problems in which Arm U was bad, which have no QC stage. Panel B: linear probability models; the three definitions of over-switchers are described in the text and classify 175, 147 and 135 subjects; rows 5 and 6 use the over-switchers of each definition who reach the QC stage. Full population: all 300 subjects; rows 5 and 6 on 266 and 267 subjects. Problem 5 alone: QU-stage over-switchers defined by their QU-stage behavior in Problem 5 and QC-stage over-switchers by their QC-stage behavior in it (42 subjects); switching back to Arm U uses all QU-stage Arm F decisions. Returning waves: only the Arm F waves after which the subject chooses Arm U again in the same problem, which drops the last wave of each problem when it is on Arm F (5,729 decisions); row 1, +1.85 ($p = 0.464$); randomization p for row 2, 0.022; rows marked – do not apply.

full population, every Arm F estimate has the same sign and is significant at the 10% level, with smaller coefficients, as expected from pooling over-switchers with subjects who rarely switch away from Arm F. In Problem 5, the effect of a single reward on switching back to Arm U and the association with current-wave rewards have the same sign and similar or larger size, but in a single problem only the association with current-wave rewards is significant at the 5% level ($p = 0.025$). In the QC stage, both responses to Arm U rewards are significant (rows 5 and 6). Rows 3 and 4 cannot be estimated with useful precision on one problem, because the state cells are thin and rewards in earlier waves are rare.

Over-switchers and other subjects

Table 16 repeats the Arm F estimates of Table 1 for the 148 subjects who are not over-switchers and the 122 single-switchers among them. If every subject responded to Arm F rewards in the same way, the response could not explain why only some subjects over-switch. Each group has its own fixed effects, since the groups switch at very different rates (in 24.6% of Arm F decisions for over-switchers, against 4.3% and 4.9%). Over-switchers respond about five times more at the same position in the wave (row 2) and ten times more when the round and the failures on Arm U are also held fixed (row 4), against either group, and the difference comes from two or more rewards, not from a single reward. When decisions are also matched on total earnings (row 3), the differences keep their sign but are not significant, with standard errors about twice as large.

As a check on this comparison alone, I re-estimate row 2 for each group using only the problems in which the subject over-switches, that is, switches back to Arm U before it paid; every other estimate in this appendix uses all five problems. 96 of the 148 subjects who are not over-switchers have at least one such problem, and the randomization tests remain exact because the selection depends only on choices. In these problems they respond about as much as over-switchers do in theirs (-3.03 against -4.06 percentage points; difference $p = 0.53$; single-switchers -2.91, difference $p = 0.49$). The groups therefore differ in how often they over-switch, but when they do, they respond to Arm F rewards in a similar way.

Estimated subject by subject, the responses are too imprecise to classify individuals: the median over-switcher has 46 Arm F decisions and 4 rewards, and the medians of the two groups do not differ significantly (rank-sum $p = 0.29$).

Table 16: Responses to Arm F rewards in the QU stage, by group

	Over-switchers	Not over-switchers	Difference	Single-switchers	Difference
(2) Rewards in the current wave	−3.10	−0.58	−2.52	−0.68	−2.42
	[0.002]	[0.011]	[0.013]	[0.039]	[0.021]
(3a) State held fixed	−2.79	−0.33	−2.46	−0.58	−2.22
	[0.050]	[0.773]	[0.179]	[0.654]	[0.249]
(3b) State held fixed, subject	−3.47	−1.07	−2.40	−1.33	−2.14
	[0.034]	[0.452]	[0.269]	[0.421]	[0.357]
(4) Rewards in the current wave,	−3.15	−0.32	−2.83	−0.27	−2.88
	[0.004]	[0.190]	[0.011]	[0.454]	[0.013]
Arm U failures held fixed					
Indicators for the number of rewards in the current wave					
One reward	+0.08	−1.34	+1.42	−1.55	+1.63
	[0.959]	[0.049]	[0.415]	[0.055]	[0.362]
Two or more rewards	−10.01	−1.40	−8.61	−1.31	−8.70
	[<0.001]	[0.040]	[<0.001]	[0.143]	[<0.001]
Switching from Arm F (%)	24.6	4.3		4.9	
Decisions	7,249	10,165		7,502	
Subjects	147	148		122	

Notes: Linear probability models, coefficients in percentage points, p-values in brackets, standard errors clustered by subject. Arm F decisions in the QU stage, all five problems. Controls as in Table 1, estimated separately for each group. Not over-switchers: 122 single-switchers and 26 subjects who did not try Arm U in their last problem in which Arm U never paid. Indicators: one indicator for one reward and one for two or more rewards in the current wave, with the controls of row 2.

Table 17: Timing of the response to a single Arm F reward, QU stage

	Mean after a failure	Effect of a reward	Rand. p	Holm p	Decisions
Panel A. Outcomes of a single draw					
Switches from Arm F in the next round (%)	24.6	−0.40 (2.08) [0.849]			7,249
Rounds stayed on Arm F	3.85	+0.46 (0.22) [0.039]	0.006		7,249
Switches back to Arm U before the end (%)	78.5	−4.58 (1.48) [0.002]	<0.001		6,629
Draw in rounds 1–10	97.2	+0.10 (1.36) [0.943]			1,899
Draw in rounds 11–20	85.6	−3.24 (2.18) [0.137]			2,629
Draw in rounds 21–28	52.5	−7.75 (2.84) [0.006]			2,101
Panel B. Has switched from Arm F within ... (%), decisions with at least eight rounds remaining					
1 round	26.5	+0.19 (2.29) [0.935]		1.000	5,652
2 rounds	41.6	−1.04 (2.36) [0.659]		1.000	5,652
3 rounds	52.6	−3.70 (2.50) [0.140]		0.427	5,652
4 rounds	60.8	−3.95 (2.45) [0.107]		0.427	5,652
5 rounds	67.1	−4.16 (2.34) [0.076]		0.401	5,652
6 rounds	71.7	−4.02 (2.19) [0.067]		0.401	5,652
7 rounds	75.8	−4.81 (2.03) [0.018]		0.124	5,652
8 rounds	78.9	−5.28 (2.04) [0.010]		0.078	5,652

Notes: Linear probability models; QU-stage over-switchers, Arm F decisions in the QU stage, all five problems. The regressor is the outcome of the round just played (reward = 1). Controls: position in the wave interacted with the number of failures since the last reward in the wave, and subject fixed effects; for rounds stayed and switching back to Arm U, also the number of rounds remaining in the problem. Rounds stayed: rounds on Arm F after the draw until the subject switches away from it or the problem ends. Switching back to Arm U: equals one if the subject chooses Arm U in any later round of the problem; decisions after the subject has tried Arm U in the problem. Coefficients in percentage points (rounds for rounds stayed); standard errors clustered by subject in parentheses, p-values in brackets. Rand. p: 5,000 redraws of every QU-stage Arm F outcome, choices held fixed, recomputing the regressor and the failures since the last reward; switching back, 3 of 5,000 redraws as extreme as the observed coefficient. Equality of the effect on switching back across the three periods of the draw: $p = 0.043$, from one regression with separate coefficients and controls for each period. Holm p: adjusted across the eight horizons of Panel B. Logit and Arm U bad estimates for switching back: Table 15; Arm U bad, rounds stayed: +0.49 ($p = 0.030$).

Timing of the response to a single Arm F reward

Table 17 follows a single Arm F reward beyond the next choice. Once a subject is on Arm F, the outcome of a round is a random draw, so its effect on later choices is causal; a reward has no effect on the next choice (row 1 of Table 1), but it could still matter over longer horizons. Panel A reports three outcomes of the same draw: switching in the next round, the number of rounds the subject stays on Arm F, and switching back to Arm U before the end of the problem. A reward lengthens the stay slightly and lowers the probability of switching back to Arm U before the end of the problem. This last effect comes from draws late in the problem (+0.10 percentage points for draws in rounds 1 to 10 and −7.75 for draws in rounds 21 to 28; equality across the three periods, $p = 0.043$), and does not vary with the position in the wave ($p = 0.76$) or with the failures since the last reward ($p > 0.99$).

Panel B follows the draw round by round on the decisions with at least eight rounds remaining, so that every horizon uses the same decisions. No horizon is significant once the p-values are adjusted for the eight horizons with Holm's method (smallest adjusted $p = 0.078$).

The number of rewards in the current wave

Table 18 adds one check to the main text: separating two rewards from three or more (Panel B), the response keeps growing with a third reward, and two rewards have more than twice the effect of one ($p = 0.001$).

Separating, in one regression with the controls of row 4, the reward of the round just played from the earlier rewards of the current wave gives the same picture: earlier rewards are associated with less switching, while the reward just played has no detectable effect (Panel D; −0.37, $p = 0.853$). A rule that conditions only on the last outcome, such as win-stay/lose-shift, implies the opposite.

The order of a reward in the wave. Panel C adds to the regression of row 1 indicators for one, two, and three or more earlier rewards in the wave and their interactions with the reward of the round just played. Each estimate compares a reward with a failure in the same round, after the same number of earlier rewards, and is causal for the subjects still on Arm F at that point of the wave. A first reward of the wave has no detectable effect, while a second and a third lower switching in the next round; the 21 fourth or later rewards are too few for a precise estimate.

Table 18: The number of Arm F rewards in the current wave, QU stage

	All problems	Arm U bad	Returning waves
Panel A. Current wave, one reward and two or more (controls of row 2)			
One reward	+0.08	−0.50	+1.91
	[0.959]	[0.768]	[0.321]
Two or more rewards	−10.01	−11.00	−12.64
	[<0.001]	[<0.001]	[<0.001]
Randomization p: two or more	<0.001	<0.001	<0.001
Panel B. Current wave, two and three or more separately (controls of row 2)			
One reward	+0.06	−0.57	
	[0.972]	[0.738]	
Two rewards	−8.66	−9.50	
	[<0.001]	[<0.001]	
Three or more	−13.52	−15.60	
	[<0.001]	[<0.001]	
Randomization p: two rewards	<0.001	<0.001	
three or more	<0.001	0.001	
Two rewards = twice one reward, p	0.001	0.011	
Panel C. Reward just played, by its order in the wave (controls of row 1)			
First reward of the wave	+2.41	+2.14	+4.68
	[0.320]	[0.414]	[0.091]
Second	−10.44	−9.39	−13.84
	[0.001]	[0.025]	[0.006]
Third	−14.14	−15.08	−15.00
	[<0.001]	[0.002]	[0.018]
Fourth or later	−1.90	+9.67	
	[0.810]	[0.365]	
Randomization p: second	0.005	0.031	0.006
third	0.045	0.063	0.184
Panel D. Reward just played and earlier rewards of the wave (controls of row 4)			
Earlier rewards of the current wave	−3.95	−4.37	
	[<0.001]	[<0.001]	
Reward in the round just played	−0.37	−0.03	
	[0.853]	[0.987]	
Randomization p: earlier rewards	<0.001	<0.001	
Decisions	7,249	5,537	5,729
Subjects	147	145	147

Notes: Linear probability models; QU-stage over-switchers, Arm F decisions in the QU stage. Coefficients in percentage points, relative to no reward; p-values in brackets, standard errors clustered by subject. Controls: those of row 2 of Table 1 (Panels A and B), row 1 (Panel C) and row 4 (Panel D). Panels A and B: indicators for the number of rewards in the current wave (1,598, 461 and 272 decisions with one, two, and three or more, all problems). Panel C: indicators for one, two, and three or more earlier rewards in the wave and their interactions with the reward of the round just played, so that each coefficient compares a reward with a failure in the same round (536, 113, 43 and 21 rewards that are the first, second, third, and fourth or later of the wave). Panel D: the rewards of the current wave split into the reward of the round just played and the earlier rewards of the wave; logit average marginal effects, −4.95 ($p < 0.001$) and −0.44 ($p = 0.824$). Returning waves: only the Arm F waves after which the subject chooses Arm U again in the same problem. Randomization p: 5,000 redraws of every QU-stage Arm F outcome, choices held fixed, recomputing every regressor; "<0.001": at most 4 of 5,000 redraws as extreme as the observed coefficient.

Keeping only the waves that return to Arm U

Because the last wave of a problem, when it is on Arm F, by definition does not end with a switch back to Arm U and can accumulate rewards, the "Returning waves" columns of Tables 15 and 18 repeat the Arm F results on the subset of Arm F waves after which the subject switches back to Arm U. This drops 1,520 of the 7,249 decisions (21.0%), which come late in the problem. Waves with several rewards often end with a switch back: 58% of those that reach a second reward (66 of 113) and 42% of those that reach a third (18 of 43). The estimates in this sample are correlational, because whether a wave is the last one depends on later outcomes and choices. The randomization test remains valid, because if rewards did not affect choices, which waves return would be the same under any draw of the rewards.

The main results hold: a single Arm F reward has no detectable effect on the next choice (row 1, +1.85, $p = 0.464$); two or more rewards in the current wave are associated with a much lower probability of switching (-12.64 percentage points, $p < 0.001$; randomization $p < 0.001$); a second reward lowers the probability of switching in the next round (-13.84, $p = 0.006$; randomization $p = 0.006$); and current-wave rewards count more than earlier-wave ones (row 4, difference -3.69, $p = 0.039$).

The association with current-wave rewards (rows 2 and 3) and the effect of a third reward keep their size but are less precise (p up to 0.242).

Current and earlier waves: rewards of the same age

Panel A of Table 19 breaks down by age the main-text comparison of rewards of the same age in the current and earlier waves, which cannot fully separate a recency effect from attention limited to the current wave. Current-wave rewards are associated with less switching at every age up to ten rounds and earlier-wave rewards at none, but no gap is significant at the 5% level. The overall estimate (-3.27, $p = 0.100$) has logit $p = 0.049$, and is -4.03 when Arm U is bad ($p = 0.085$).

Age in Arm F draws. Measured in Arm F draws, where recent earlier-wave rewards are more common (127 to 222 lie two to five draws back, against 272 to 524 current-wave rewards), the gap is -5.17 ($p = 0.003$; randomization $p < 0.001$; -5.76 when Arm U is bad, $p = 0.005$; -4.30 with five coarser groups, $p = 0.004$).

The controls of row 4. Before Arm U pays, the round and the failures on Arm U are, with earnings (row 3), the state of the problem: they fix the posterior probability that Arm U is good and the rounds left. Holding the earlier-wave Arm F rounds fixed directly instead, with the position in the wave and subject fixed effects, gives the same gap between current

and earlier waves (-4.08, $p = 0.006$).

Table 19: Current and earlier waves: rewards of the same age

Rounds before the decision	Current wave	Earlier waves	Difference
Panel A. QU stage, Arm F (QU-stage over-switchers)			
3–5	−4.57 [<0.001]	+0.99 [0.735]	−5.56 [0.077]
	(1,027)	(264)	
6–10	−3.93 [0.004]	−0.44 [0.781]	−3.49 [0.104]
	(791)	(775)	
11 or more	−1.92 [0.436]	+0.79 [0.523]	−2.72 [0.348]
	(405)	(1,713)	
Panel B. QC stage, Arm U (QC-stage over-switchers)			
All ages	−3.16 [<0.001]	−0.45 [0.763]	−2.71 [0.097]
Same age: additional effect of the current wave			−0.72 [0.675]

Notes: Linear probability models, all five problems. Panel A: one regression; QU-stage over-switchers, Arm F decisions in the QU stage (7,249 decisions). Coefficients in percentage points per reward, p-values in brackets, standard errors clustered by subject; number of rewards in parentheses. Controls: those of row 4 of Table 1, Arm F rewards one or two rounds before the decision, and the number of earlier-wave Arm F rounds in each age bin. Estimate in the text: the controls of row 4, past Arm F rewards in nine age bins (1, 2, 3, 4, 5, 6–7, 8–10, 11–15 and 16 or more rounds before the decision, in any wave) and earlier-wave Arm F rounds in each bin of three or more rounds. These are held fixed because an earlier-wave Arm F round three to five rounds back marks a subject who switches back and forth quickly (in 40.3% of decisions, against 19.2%); without them, earlier-wave rewards at that age appear to raise switching ($+5.72$, $p = 0.037$). Arm F draws: one indicator for each number of Arm F draws back, in the same bins, and the controls of row 4. Panel B: QC-stage Arm U decisions (3,640 decisions, 103 subjects); all ages, the controls of row 6 of Table 1 plus the number of post-payment Arm U rounds in earlier waves; same age, the same plus the time since Arm U paid, post-payment Arm U rewards in nine age bins in any wave, and earlier-wave post-payment Arm U rounds in each age bin; logit average marginal effect of the difference between current and earlier waves, -4.05 ($p = 0.009$).

Settling on Arm F when Arm U is bad

Table 20 compares how often over-switchers and single-switchers stop choosing Arm U for the rest of the problem when Arm U was bad. A subject who learns from Arm U's failures should eventually settle on Arm F. The comparison excludes the problem used to classify the subject, in which the two groups are defined by whether they switched back to Arm U, so their settling differs by definition. Over-switchers settle far less often at every round.

Table 20: Settling on Arm F when Arm U is bad

	Over-switchers	Single-switchers	Difference
Settled on Arm F by round (% of problems)			
10	4.1	29.4	−25.4
15	9.9	51.5	−41.6
20	15.3	62.3	−46.9
25	29.7	71.1	−41.3
Share of rounds on Arm U after the first move to Arm F (%)			
	35.7	12.7	+23.0
Problems	222	204	
Subjects	119	111	

Notes: Problems in which Arm U was bad and the subject chose it at least once, excluding the problem used to classify the subject. A problem is settled on Arm F by round r if the subject does not choose Arm U in rounds r to 30. Share of rounds on Arm U after the first move to Arm F: share of the rounds after the subject's first move from Arm U to Arm F that are spent on Arm U (220 and 202 problems with such a move before round 30). Single-switchers tried Arm U in their last problem in which Arm U never paid but did not return to it after switching to Arm F. Differences in percentage points from linear probability models with standard errors clustered by subject; every difference has $p < 0.001$.

The QC stage

In the QC stage,[43] the gap between current and earlier waves on Arm U has logit $p = 0.009$ ($p = 0.10$ with round fixed effects), but rewards of the same age show no additional effect of the current wave (Table 19, Panel B). On Arm F (Table 21), I detect no response to rewards on the next choice, though with only 153 rewards the standard errors are 2.6 to 4.1 times those of the QU stage; over longer horizons, a reward raises switching back to Arm U, the opposite of the QU stage (difference 9.79 percentage points, $p < 0.001$).

Table 21: Arm F in the QC stage, QC-stage over-switchers

	Estimate
Same specifications as in the QU stage	
(1) Reward in the round just played	+2.18 (5.42) [0.687]
(2) Rewards in the current wave	+0.70 (4.12) [0.865]
(4) Rewards in the current wave	+1.37 (4.08) [0.737]
Rewards in earlier waves	+2.08 (2.26) [0.358]
Difference	−0.71 (4.49) [0.874]
One reward in the current wave	+2.96 (5.01) [0.555]
Two or more rewards	−1.95 (9.09) [0.830]
Reward just played → rounds stayed on Arm F	−0.52 (0.27) [0.052]
Reward just played → switches back to Arm U	+5.22 (1.78) [0.003]
Decisions / rewards	1,519 / 153
Subjects	102

Notes: Linear probability models; QC-stage over-switchers, Arm F decisions in the QC stage, all five problems. Coefficients in percentage points (rounds for rounds stayed), standard errors clustered by subject in parentheses, p-values in brackets. The regressions of rows 1, 2 and 4 of Table 1, of Panel A of Table 18 (indicators for one reward and for two or more rewards in the current wave) and of Table 17 (rounds stayed and switching back), with the same controls, on the Arm F decisions of the QC stage; decisions with one, two, and three or more rewards in the current wave: 280, 46 and 6. Difference between stages: one regression stacking the QU-stage decisions of QU-stage over-switchers and the QC-stage decisions of QC-stage over-switchers, with every fixed effect specific to the stage.

Arm U before and after it paid. Do over-switchers switch away from Arm U less once it has paid? Yes: the 83 subjects who are over-switchers in both stages switch away from it in 21.7% of decisions after it paid against 34.3% before, 13.3 percentage points less for the same subject in the same round ($p < 0.001$). But the rounds right after the payment

[43] Two subjects whose Arm U first paid in round 29 and who stayed on it in round 30 have no other QC-stage decision and are counted as not over-switching. Counting them as over-switchers, the effect of an Arm U reward in the round just played on switching away from Arm U (row 5) is −5.34 percentage points ($p = 0.020$), and the association with rewards in the current Arm U wave (row 6) is −3.02 ($p < 0.001$).

follow a reward, so I compare failures only: a failure in a later Arm U wave, before any further reward, is followed by a 6.63 percentage point lower probability of switching than a failure at the same position of a wave before Arm U paid, for the same subject and round ($p = 0.008$; 78 subjects; 4,290 failures). When I restrict the comparison further to failures within the same problem, before and after Arm U paid (152 problems), the observations fall from 4,290 to 1,253 failures and the difference is not significant (-0.94, $p = 0.87$; with subject-by-problem fixed effects, -0.23, $p = 0.97$, 95% confidence interval -13.2 to $+12.7$).

Models of switching

This part gives the samples, definitions and estimates behind Table 22 that are not reported in the main text. Decisions follow the outcomes of rounds 1 to 28 in all five problems, including the problem used to classify the subject.

Threshold rules. Under a rule that switches away from an arm after K consecutive failures, with the count reset by a reward, a subject switches right after a reward only by mistake, and every Arm U wave before it pays, in which every draw is a failure, lasts K rounds. QU-stage over-switchers switch away from Arm F in 23.8% of decisions right after a reward and in 24.6% after a failure (row 1 of Table 1 gives the causal comparison). Among the 143 QU-stage over-switchers with three or more Arm U waves that ended with a switch to Arm F before Arm U paid, 71% have one wave that ended after a single failure and another that lasted five rounds or more; the median difference between a subject's longest and shortest wave is 6 rounds. Both implications concern a single arm, so they hold whether or not the threshold differs between arms. A rule under which the probability of switching rises with the failures since the last reward, rather than jumping at K, still implies that only those failures matter. With one indicator for each combination of position in the wave and failures since the last reward (after the round just played) and subject fixed effects, each earlier reward in the wave is associated with a 7.25 percentage point lower probability of switching away from Arm F ($p < 0.001$; logit $p < 0.001$), and 6.98 when Arm U is bad ($p < 0.001$).

Hot-hand beliefs. Among QU-stage over-switchers, 88% of switches occur in waves without a reward so far: all 1,882 switches away from Arm U, which has not paid in the QU stage, and 1,355 of the 1,781 switches away from Arm F (76%); when Arm U is bad, 89%. With indicators for one reward and for two or more rewards in the current wave, the second split by whether the last two rewards are consecutive, and position in the wave and subject fixed effects, two or more rewards are associated with a probability of switching away from

Table 22: Models of switching: predictions and results

Model	Prediction	Result	Verdict
Panel A. Choices respond to outcomes that carry no information			
Win-stay/lose-shift	Only the last outcome matters	A single Arm F reward has no detectable effect (-0.40, $p = 0.849$; row 1 of Table 1); each earlier reward in the wave, last outcome held fixed: -3.95 ($p < 0.001$)	Rejected
Switch after K consecutive failures (K may differ by arm)	Almost no switching right after a reward; waves of equal length; only the failures since the last reward matter	Switching from Arm F right after a reward as frequent as after a failure (row 1 of Table 1); for 71% of over-switchers, one Arm U wave ends after one failure and another lasts five rounds or more; each earlier reward in the wave, same position and failures since the last reward: -7.25 ($p < 0.001$)	Rejected
Reinforcement learning, values carried across waves	At equal Arm F draws back, current- and earlier-wave rewards count equally	Current-wave reward relative to an earlier-wave reward at equal draws back: -5.17 ($p = 0.003$)	Rejected
Arm F of unknown but fixed quality	As for Arm U, a single reward reveals a good arm; with an unknown reward probability, every reward counts, whatever its wave	Single reward: -0.40 (row 1 of Table 1); earlier-wave rewards $+0.74$, against -3.15 in the current wave (row 4 of Table 1)	Rejected
Quality changing from wave to wave, learnable within the wave	Only current-wave rewards lower switching; waves with several rewards rarely end	Current wave -3.15, earlier waves $+0.74$, difference -3.88 ($p = 0.013$; row 4 of Table 1); 58% of waves that reach a second reward, and 42% of those that reach a third, end with a return to Arm U	Rejected
Runs of failures (cold streak)	Switching rises with the failures since the last reward	Per failure: Arm F, QU stage, $+2.02$ ($p = 0.007$); Arm U, QC stage, $+2.68$ ($p < 0.001$)	Consistent, small
Panel B. Choices do not respond to outcomes			
Random switching, rate per arm constant or changing over time	Outcomes do not affect choices	Constant rate: switching from Arm F falls by 7.47 per ten rounds ($p < 0.001$). Rate changing over time: randomization $p \leq 0.001$ (rows 2, 4 and 6 of Table 1); Arm U reward in the QC stage: -6.02 ($p = 0.007$; row 5 of Table 1)	Rejected
Probability matching	Arm U chosen with probability $\bar{\theta}_U/(\bar{\theta}_U + 0.1)$, independently of the last choice; QC stage: two thirds	QC stage: 47% of subjects choose Arm U more than 5 points above two thirds, 35% more than 5 points below, 18% within 5 points (53% if all matched); moves from Arm U to Arm F in 20.4% of decisions and back in 44.1%, against 33.3% and 66.7% ($p < 0.001$). QU stage: Arm U $+15.09$ relative to matching ($p < 0.001$); 76% of subjects more than 5 points above matching, 10% more than 5 points below, 14% within 5 points (71% if all matched)	Rejected
Thompson sampling	Each arm chosen with	55% of subjects with a problem in	Rejected

Arm F 9.64 percentage points lower when the last two are consecutive (258 decisions) and 10.24 lower when they are not (475 decisions; $p < 0.001$ for both). The difference is +0.60 ($p = 0.80$; logit $p = 0.90$; when Arm U is bad, +2.40, $p = 0.45$).

Runs of failures. For Arm F in the QU stage (QU-stage over-switchers) and Arm U in the QC stage (QC-stage over-switchers), the run is the number of draws since the last reward in the current wave; on Arm U in the QC stage the wave is counted from Arm U's first payment, which counts as a reward. The sample is the decisions that follow a failure in waves with an earlier reward: 1,618 decisions of 125 subjects on Arm F and 1,718 decisions of 101 subjects on Arm U. The regression holds fixed the subject, the round, the position in the wave and the number of rewards in the wave (one, two, three or more), with the run capped at six, so the length of the run varies only with when the rewards came. Each additional failure is associated with a 2.02 percentage point higher probability of switching on Arm F ($p = 0.007$; when Arm U is bad, 2.59, $p = 0.012$) and by 2.68 on Arm U ($p < 0.001$).

Random switching. If the probability of switching away from each arm does not depend on outcomes, choices do not depend on outcomes, which the randomization tests of Table 1 reject. The probability of switching is not constant either. For QU-stage over-switchers, in waves with no reward so far (all 5,887 decisions on Arm U before it pays and 4,918 decisions on Arm F before the first reward of the wave), holding fixed the subject and the position in the wave, the probability of switching away from Arm F falls by 7.47 percentage points per ten rounds ($p < 0.001$), and the probability of switching away from Arm U rises by 2.63 ($p = 0.009$). In the QC stage, the share of Arm U choices of QC-stage over-switchers falls by 2.57 percentage points per ten rounds after Arm U's first payment, for the same subject ($p = 0.044$).

Probability matching. Under probability matching, a subject chooses Arm U with probability $\bar{\theta}_U/(\bar{\theta}_U + 0.1)$, where $\bar{\theta}_U$ is the expected reward probability of Arm U, independently of the arm chosen before. In the QC stage $\bar{\theta}_U = 0.2$, so the probability is two thirds. In the QU stage, after k failures on Arm U, $\bar{\theta}_U = 0.2\,\pi_k$, where $\pi_k = 0.8^k/(0.8^k + 1)$ is the posterior probability that Arm U is good, so the probability falls from one half before any draw to 14% after eleven failures. In the QC stage (all choices after Arm U's first payment, in all problems in which Arm U paid), QC-stage over-switchers choose Arm U in 67.4% of their choices on average across subjects ($p = 0.72$ against two thirds; 69.1% pooled over choices), but individually 47% choose it more than five percentage points above two thirds, 35% more than five points below, and 18% within five points. The other 164 subjects with a problem in which Arm U paid choose Arm U in 97.2% of their choices, and 74% of them never choose

Arm F. Choices persist: QC-stage over-switchers switch from Arm U to Arm F in 20.4% of decisions and from Arm F to Arm U in 44.1%, against 33.3% and 66.7% under matching ($p < 0.001$ for both). In the QU stage (all decisions before Arm U's first payment), QU-stage over-switchers choose Arm U next in 44.05% of decisions, against 28.96% under matching (+15.09 percentage points, $p < 0.001$), and the gap grows with failures: after eleven or more, 47% against 9%. Individually, 76% choose Arm U more than five percentage points above the matching share, 10% more than five points below, and 14% within five points.

To compare these individual shares with what chance alone produces if every over-switcher matched, I simulate 10,000 samples in which every over-switcher matches, keeping each subject's own number of choices: in the QC stage each choice is Arm U with probability two thirds, and in the QU stage each decision chooses Arm U next with the matching probability at the subject's number of Arm U failures. A matcher's share differs from the matching share by chance, more so with fewer choices (with the median of 51 QC-stage choices, its standard deviation is 6.6 percentage points). Chance places on average 53.1% of subjects within five points of the matching share in the QC stage (95% of simulated samples between 43.7% and 62.1%) and 71.4% in the QU stage (63.9% to 78.2%), against 18.4% and 13.6% observed; it places 23.7% above and 23.2% below in the QC stage, against 46.6% and 35.0% observed, and 14.3% on each side in the QU stage, against 76.2% and 10.2%. No simulated sample has as few subjects within five points or as many above as observed, in either stage ($p < 0.001$); in the QC stage, 0.5% have as many below ($p = 0.005$), while in the QU stage the share below is within what chance produces. For the same subject at the same number of Arm U failures, being on Arm U is associated with a 36.62 percentage point higher probability of choosing it next ($p < 0.001$). The other 148 classified subjects choose Arm U next in 29.59% of decisions, against 32.09% under matching ($p = 0.22$), with the same persistence (+71.04 percentage points, $p < 0.001$); the 122 single-switchers among them choose Arm U next in 32.90% of decisions, against 29.43% under matching ($p = 0.051$; when Arm U is bad, 31.12% against 27.40%, $p = 0.038$), with persistence of +69.51 percentage points ($p < 0.001$).

Thompson sampling. Thompson sampling chooses each arm with the probability that it is the better one. With correct beliefs, Arm U is known to be better once it has paid, so Arm F is never chosen in the QC stage; 146 of the 267 subjects with a problem in which Arm U paid (55%) choose Arm F at least once in that stage, and Arm F accounts for 13.6% of QC-stage choices. If subjects are instead uncertain about Arm U's reward probability, with a uniform prior updated with every Arm U draw in the problem, including the failures before Arm U

paid, Arm F is chosen with the posterior probability that Arm U's reward probability is below 0.1. For QC-stage over-switchers this implies Arm F in 11.58% of QC-stage decisions, against 30.84% observed. Their choices also depend on more than the numbers of Arm U rewards and failures: holding these numbers and the subject fixed, QC-stage over-switchers are 21.34 percentage points more likely to choose Arm F next when they are already on it ($p < 0.001$), and on Arm U a reward in the round just played lowers the probability of choosing Arm F next by 6.52 percentage points ($p = 0.001$).

Rewards per wave. Of the 2,002 Arm F waves of QU-stage over-switchers in the QU stage, 73% contain no reward up to the last decision of the wave, 21% contain one and 6% contain two or more.

7.7 Comparison with the probability-matching literature

Probability matching is the tendency to choose each of two options in proportion to its probability of paying off, instead of always choosing the more likely one (Vulkan, 2000). I assume Bayesian updating, so matching means choosing Arm U with probability $\bar{\theta}_U/(\bar{\theta}_U+0.1)$, where $\bar{\theta}_U$ is Arm U's expected reward probability: 0.2 in the QC stage, so that the matching share is two thirds, and $0.2\,\pi_k$ after k failures in the QU stage, with $\pi_k = 0.8^k/(0.8^k+1)$. Table 23 compares the choices in my data with the literature. In the QC stage subjects know both reward rates, so the benchmark is studies that tell subjects the probabilities (panel A). In the QU stage they learn Arm U's quality from its failures, so the benchmarks are studies that do not tell them (panel B) or tell them only in part (panel C).[44]

In the QC stage, 45% of the 267 subjects with a problem in which Arm U paid never choose Arm F after Arm U's first payment, 53.6% choose Arm U in more than 95% of their choices, and the mean share of Arm U is 85.7%, 19 points above matching. The closest design is Newell and Rakow (2007): subjects know that one outcome has probability two thirds and predict 300 rolls with feedback. Under the same 95% criterion, their maximizers rise from 28% in the first 60 predictions to 56% in the last 60, and the mean share from .77

[44] In every matching study in panels A and B, subjects bet on which of several mutually exclusive outcomes of one event occurs; I found none in which two options pay independently, as the arms here do. With mutually exclusive outcomes, the less likely option pays exactly when the likely one fails, which gives mixing an appeal as a hedge that independent options lack; with feedback, maximization is 85% when two payoffs are mutually exclusive and 97% when they are positively correlated (Erev et al., 2017). I match reward rates, as Meyer and Shi (1995, fn. 6) do; matching the probability that each arm is the only one to pay gives 0.69 in the QC stage, where 17.5% of QC-stage over-switchers are within five points of it (54% if all matched, $p < 0.001$), and lowers the matching share of Arm U in the QU stage by at most 1.3 points, leaving the comparisons unchanged.

to about .91.[45] The subjects in this experiment lie on this path and, like theirs, move toward the better option with experience: among the 197 subjects with two or more problems in which Arm U paid, the share above 95% rises from 51% in the first to 66% in the last ($p < 0.001$). Non-maximizers choose Arm U in 69.7% of choices under the 95% criterion (73.8% under the strict one), at or just above non-maximizers in tasks without feedback (-4 to $+3$ points) and below those in the last game of Newell and Rakow ($+13$ to $+19$). QC-stage over-switchers average 67.4%, at matching. Outside the problem used to classify them they also move toward Arm U across problems (from 65.4% to 71.2%, $p = 0.044$), but within a problem their share of Arm U falls by 2.6 points per ten rounds after Arm U's first payment. Only 18% are within five points of two thirds, against 53% if all of them matched, and 35% are more than five points below, against 6–33% of non-maximizers 10 or more points below matching in stated-probability studies (West and Stanovich, 2003; Barlow et al., 2025).

In the QU stage I use the decisions in which both Thompson sampling and the optimal policy choose Arm F; matching then implies Arm F in 81% of them.[46] All 288 subjects choose Arm F in 70.3% of these decisions, 11 points below matching, and 53% are more than five points below. QU-stage over-switchers choose it in 60.9%, against 82.3% under matching, and 74% are more than five points below. The other 141 subjects average the matching share (80.0% against 79.5%), with half of them more than five points above it. In studies that do not tell subjects the probabilities, final-block group means of the more likely option lie between 4 points below and 25 points above matching, every paid study with symmetric payoffs in Vulkan's (2000) tables is above it, and where individual data are reported, 13–21% of subjects are at least five points below (Gaissmaier and Schooler, 2008).[47] Over-switchers

[45] Their payments are hypothetical, but a follow-up with real payments "did not differ significantly" (Newell and Rakow, 2007, fn. 2), and matching forgoes one sixth of the maximum expected payoff in both designs. Without feedback, 9–13% of subjects maximize in sequential tasks of 50 to 300 predictions (Newell et al., 2013) and 15–47% in sets of 5 to 10 predictions (e.g., Koehler and James, 2010; West and Stanovich, 2003), with mean shares 4 to 15 points above matching. In the two paid economics experiments I found, 75.5% always choose the likely option (Kim and Kim, 2022) and 36% choose the same bet in at least 90% of 20 repetitions (Agranov, Healy and Nielsen, 2023). My strict criterion covers every QC-stage choice of every problem in which Arm U paid (a median of 53); in the last problem in which Arm U paid alone, 63.7% choose Arm U in more than 95% of their choices.

[46] Subjects see only the outcome of the arm they choose, so choosing Arm U carries information; the optimal policy prices it, and I keep only decisions in which it still chooses Arm F. Thompson sampling, also called probability matching in the bandit literature, chooses Arm F with probability $1 - \pi_k$, above one half from the first failure on. I do not elicit beliefs: at the average shares, over-switchers' choices correspond to matching a probability of about .32 that Arm U is good, against an average posterior of about .11. Over-switchers are classified on switching, which moves their shares toward one half; the full-sample figures are free of this.

[47] In the first 50 trials, before the probabilities are learned, group means are at or below matching: .60 and .62 at $p = .70$ (McMahon and Scheel, 2010) and .70 (Ert and Bereby-Meyer, reported by Erev and Haruvy,

thus deviate toward the option with the lower expected reward, whereas group means in this literature deviate toward the more likely option once the probabilities are learned. Among bandit experiments, the closest design is Meyer and Shi (1995): an airline with a known on-time rate against an ambiguous one, paid per on-time departure. When the optimal policy prescribes no trips on the ambiguous airline, subjects take it on 20 to 26 percentage points more trips than that policy, between matching (44–49%) and maximizing. The other subjects choose Arm U in 20.0% of the decisions in which Arm F is optimal, at matching, and over-switchers in 39.1%, more than twice the 17.7% that matching implies.[48]

7.8 Formal Models and Propositions

Baseline One-armed Model: Environment and Preferences

A decision maker (DM) faces a finite-horizon one-armed bandit problem with horizon $N \in \mathbb{N}$. In each round $t \in \{1, \ldots, N\}$, the DM chooses an action $x_t \in \{U, F\}$.

Arms. Arm U has an unknown type. With prior probability $p \in (0, 1)$, Arm U is good. If good, it yields a reward $w > 0$ with probability $\theta_U \in (0, 1)$ in each round. If bad, it yields zero reward with probability one. Arm F has a known type and yields a reward $w > 0$ with probability $\theta_F \in (0, 1)$ in each round. I assume $\theta_U > \theta_F$ and

$$p\theta_U = \theta_F,$$

so that the two arms have ex ante identical probability of returning a reward in the first round.

Information. Rewards are observed immediately after each choice. Conditional on the type of each arm, rewards are independent across rounds. Only outcomes from Arm U convey information about the underlying state; draws from Arm F are uninformative.

Strategies. A (behavioral) strategy is a mapping from histories of past actions and realized rewards to current actions. Strategies may be contingent on the entire history of play.

2016). After 1,000 trials, 20% of subjects are at or below matching (Edwards, 1961).

[48] In bandits with a safe arm and a risky arm that is either good or never pays, as in the QU stage here, about 80% of games follow a cut-off rule in the control treatment of Hoelzemann and Klein (2021), and subjects explore less than predicted in Hudja and Woods (2024a). With two arms of unknown quality, subjects choose the arm with the higher true reward rate in 70–73% of rounds, against 80–82% under the optimal policy (Hudja and Woods, 2025).

Table 23: Probability matching: this paper and the literature

Study	Design	p	N	Max.	Mean	Non-max.
A. Probabilities told to subjects (benchmark for the QC stage)						
Newell & Rakow (2007), feedback	300 pred., fb., hyp.	.67	25	28/56[a]	.77/.91[a]	.69/.83[b]
Newell & Rakow (2007), no fb.	first 60 of 300, hyp.	.67	25	12[a]	.71	.67[b]
Koehler & James (2010), Exps. 1–2	10 pred., no fb., paid	.70	84/78	43/35	.78–.85/.79[b]	–/.68[b]
West & Stanovich (2003), Exp. 3	10 pred., hyp.	.70	422	32	.77–.79[b]	.66–.70[b]
Newell et al. (2013), Exps. 1–2	50 / 300 pred., paid	.70	123/100	13/9[c]	.79/.78–.87[d]	
Kim & Kim (2022)	4 choices/game, paid	.75	106	76/86[e]	.96	.75[b]
Agranov, Healy & Nielsen (2023)	20 choices/p, paid	.55–.80	84	36[f]		$\approx p$[f]
B. Probabilities not told, learned from outcomes (benchmark for the QU stage)						
Edwards (1961)	1,000 pred., unpaid	.60/.70	80	9[g]	.69–.89	.68–.86[b]
Vulkan (2000), Tables 1–2	51 group means	.60–.90			$p-.04$ to $p+.25$	
Shanks et al. (2002), Exps. 1–3	300–1,800 pred., paid	.70	16/12/12	38/67/75[h]	$> .90$[h]	$> .84$[b,h]
Schulze et al. (2019)	500 pred., paid	.75	162	32–43[i]		
Feher da Silva et al. (2017)	300 pred., last 100	.70	84		.77	
Erev & Haruvy (2016)[j]	500 choices, paid	.70	20		.93	
C. Bandit experiment, arm quality told in part (benchmark for the QU stage)						
Meyer & Shi (1995)	known vs. ambiguous		40		.74–.80[k]	
D. This paper (p is the matching share)						
QC stage, all subjects	after Arm U pays	.67	267	45/54[l]	.86	.74/.70[l]
QC stage, over-switchers	same	.67	103	0	.67	.67
QU stage, all subjects	Arm F optimal[m]	.81	288	8	.70	
QU stage, over-switchers	same	.82	147	1	.61	
QU stage, other subjects	same	.80	141	16	.80	

Notes: p: probability of the more likely outcome. Max.: percentage of subjects who always choose the better option. Mean: mean share of choices of the better option (Arm U in the QC stage, Arm F in the QU stage). Non-max.: mean share among subjects who are not maximizers. pred.: predictions; fb.: outcome feedback; hyp.: hypothetical payoffs. The online appendix lists all studies reviewed.

[a] First / last of five 60-prediction games; maximizer: more than 95% of a game. From Figures 1–2, except .77 and .71.

[b] Derived from reported means, counts or individual data; approximate.

[c] Every prediction, in the condition without hint or feedback. [d] Exp. 1, no hint; Exp. 2, final block without / with feedback.

[e] All eight choices of two games / all four choices of the binary game (derived). [f] Same bet in at least 90% of 20 repetitions, one paid; mixers choose the dominant bet about as often as p.

[g] Every one of the last 80 trials. [h] Max.: a 50-trial run, a 50-trial block, all of the final block; mean: Exps. 1–2; non-max.: Exp. 1.

[i] At least 95% in the final 100-trial block. [j] Experiment by Ert and Bereby-Meyer (Problem 6), full feedback, trials 401–500.

[k] Paid; 5- and 20-trip games in which the optimal policy prescribes no trips on the ambiguous airline: one minus the excess of such trips (their Table A2).

[l] Every QC-stage choice on Arm U / more than 95% of them. [m] QU-stage decisions in which both Thompson sampling and the optimal policy choose Arm F.

Preferences. The DM derives utility from the *total* accumulated reward over the entire horizon.

Let

$$S_N = \sum_{t=1}^{N} w_t$$

denote total earnings. Preferences are represented by a utility function

$$u : \mathbb{R}_+ \to \mathbb{R},$$

which is strictly increasing. Proposition 1 holds for any such u. Propositions 2–4 assume constant absolute risk aversion, $u(S) = -e^{-rS}/r$ with $r \neq 0$, or risk neutrality, $u(S) = S$, so that the earnings accumulated so far do not change the ranking of continuation strategies; Proposition 5 assumes risk neutrality. In all cases, utility need not be additively separable across periods, and payoffs are evaluated only at the end of the horizon.

Objective. The DM chooses a strategy to maximize expected utility

$$\mathbb{E}\big[u(S_N)\big],$$

where expectations are taken with respect to both payoff uncertainty and belief updating.

Because utility is evaluated only at the end of the horizon and need not be additively separable across time, I do not rely on standard Bellman equations. All results follow from dominance and monotonicity arguments applied to complete contingent strategies.

Proposition 1. *If Arm U is revealed to be good, then Arm U will be chosen in every subsequent round.*

Let

$$S_{K-1} = \sum_{j=1}^{K-1} w_j$$

denote cumulative rewards up to round $K-1$.

Assume that Arm U is known to be good, so it pays w with probability θ_U.

The proof proceeds in three steps following a backward induction approach.

Step 1. I show that choosing Arm U is optimal in the last round. In the last round ($K = N$), choosing Arm U gives an expected utility, $EU = \theta_U u(S_{K-1}+w)+(1-\theta_U)u(S_{K-1})$, while choosing Arm F gives $EU = \theta_F u(S_{K-1} + w) + (1 - \theta_F)u(S_{K-1})$.

Since $\theta_U > \theta_F$, Arm U FOSD Arm F.

Step 2 (Inductive step). Assume that for some arbitrary round $K < N$, Arm U is chosen in every round from $K+1$ to N. I show that Arm U is then also optimal in round K.

The expected value of choosing Arm U in round K is,

$$\theta_U \sum_{t=0}^{N-K} p_U^{\text{good}}(t, N-K)\, u(S_{K-1} + w + tw) + (1-\theta_U) \sum_{t=0}^{N-K} p_U^{\text{good}}(t, N-K)\, u(S_{K-1} + tw)\,,$$

where $p_U^{\text{good}}(t, T)$ denotes the probability that a good Arm U generates exactly t rewards in the next T rounds.

The expected value of choosing Arm F in round K is

$$\theta_F \sum_{t=0}^{N-K} p_U^{\text{good}}(t, N-K)\, u(S_{K-1} + w + tw) + (1-\theta_F) \sum_{t=0}^{N-K} p_U^{\text{good}}(t, N-K)\, u(S_{K-1} + tw)\,,$$

Since the probability of getting a reward in each round is independent, and hence independent of previous choices (since Arm U's quality is already known), the comparison boils down to $\theta_U > \theta_F$. Hence, Arm U FOSD Arm F.

Step 3. Since the result holds for $K = N$, and the inductive step in Step 2 holds for any arbitrary K given the hypothesis that Arm U is chosen from $K+1$ onward, backward induction gives the result for all rounds. ■

Proposition 2. *For any round K in which Arm U's quality is unknown (i.e., $p_K \in (0,1)$), no optimal strategy involves choosing Arm U immediately after choosing Arm F in round K.*

Proof. Consider some round K with posterior belief $p_K \in (0,1)$, so Arm U's quality is not fully known.

Let $FU\sigma$ denote the strategy that chooses Arm F in round K, Arm U in round $K+1$, and from round $K+2$ onward follows the expected utility maximizing continuation strategy σ. Let $UF\sigma$ denote an alternative strategy that chooses Arm U in round K, Arm F in round $K+1$ (conditional on U not delivering a reward in round K), and from round $K+2$ onward follows the same continuation strategy σ. Under constant absolute risk aversion, whether Arm U is optimal in round $K+1$ does not depend on Arm F's outcome in round K, so it is without loss to let $FU\sigma$ choose Arm U in round $K+1$ after either outcome.

If Arm U delivers a reward in its draw, then by *Proposition 1*, the continuation strategy

prescribes choosing Arm U in all remaining rounds:

$$\sigma\big(p_{K+2} = 1\big) = \{x_t = U\}_{t=K+2}^{N}.$$

Under $UF\sigma$ this would occur in round K (so $x_{K+1} = U$ as well), while under $FU\sigma$ this would occur in round $K + 1$.

I will show that $UF\sigma$ strictly dominates $FU\sigma$ in expectation.

Since Arm F's draws carry no information about Arm U's quality, p_{K+2} depends only on the outcome of the single Arm U draw. Hence when $w(U) = 0$, the posterior p_{K+2} is identical under both strategies regardless of $w(F)$, and since, under constant absolute risk aversion, σ depends only on p_{K+2} and the round, not on the earnings accumulated so far, it prescribes identical contingent actions from round $K+2$ onward under both $UF\sigma$ and $FU\sigma$.

Since both strategies therefore yield identical expected payoffs when $w(U) = 0$, and this event occurs with the same probability under both, it suffices to compare them only when Arm U delivers a reward.

Comparing the strategies.

Strategy $FU\sigma$: Arm U is drawn in round $K + 1$. There are two sub-cases.

- Both Arm F and Arm U deliver a reward. This occurs with probability $\theta_F \cdot p_K\theta_U$ and yields conditional expected utility:

$$\sum_{t=0}^{N-K-1} p_U^{\text{good}}(t,\, N-K-1)\; u(S_{K-1} + 2w + tw)\,.$$

- Only Arm U delivers a reward (Arm F does not). This occurs with probability $(1 - \theta_F) \cdot p_K\theta_U$ and yields conditional expected utility:

$$\sum_{t=0}^{N-K-1} p_U^{\text{good}}(t,\, N-K-1)\; u(S_{K-1} + w + tw)\,.$$

Strategy $UF\sigma$: Arm U is drawn in round K. After a U-success, *Proposition 1* prescribes choosing Arm U in all subsequent rounds. To keep payoffs comparable with $FU\sigma$, I separate the payoff from round $K + 1$ explicitly.

- Arm U delivers a reward in both rounds K and $K + 1$. This occurs with probability

$p_K\theta_U^2$ and yields conditional expected utility:

$$\sum_{t=0}^{N-K-1} p_U^{\text{good}}(t,\, N-K-1)\; u(S_{K-1}+2w+tw)\,.$$

- Arm U delivers a reward only in round K (not in $K+1$). This occurs with probability $p_K\theta_U(1-\theta_U)$ and yields conditional expected utility:

$$\sum_{t=0}^{N-K-1} p_U^{\text{good}}(t,\, N-K-1)\; u(S_{K-1}+w+tw)\,.$$

where $p_U^{\text{good}}(t,T)$ denotes the probability that a good Arm U generates exactly t rewards in T rounds. Note that the sums run over $N-K-1$ remaining rounds after the two-round comparison window.

Letting $V_2 \equiv \sum_t p_U^{\text{good}}(t, N-K-1)\, u(S_{K-1}+2w+tw)$ and $V_1 \equiv \sum_t p_U^{\text{good}}(t, N-K-1)\, u(S_{K-1}+w+tw)$ for brevity, $UF\sigma$ dominates $FU\sigma$ if:

$$p_K\theta_U^2\, V_2 + p_K\theta_U(1-\theta_U)\, V_1 \qquad > \theta_F \cdot p_K\theta_U\, V_2 + (1-\theta_F)\cdot p_K\theta_U\, V_1.$$

Dividing both sides by $p_K\theta_U > 0$:

$$\begin{aligned} &\theta_U\, V_2 + (1-\theta_U)\, V_1 \\ &\quad > \\ &\theta_F\, V_2 + (1-\theta_F)\, V_1. \end{aligned}$$

Rearranging:

$$(\theta_U-\theta_F)\, V_2 > (\theta_U-\theta_F)\, V_1.$$

Since $\theta_U > \theta_F$, this reduces to $V_2 > V_1$, which holds because u is strictly increasing and $S_{K-1}+2w+tw > S_{K-1}+w+tw$ for all t.

Since $UF\sigma$ strictly dominates $FU\sigma$, no optimal strategy prescribes choosing Arm U immediately after Arm F. ■

Proposition 3. *Optimal policy involves at most one switch (from Arm U to Arm F).*

Proof. Since Proposition 2 applies for any round $K < N$ with $p_K \in (0, 1)$, no optimal policy can prescribe a switch from Arm F to Arm U at any history in which Arm U's quality remains unknown. Hence if $x_K = F$ is optimal, it must be that $x_{K+1} = F$. Applying the same argument repeatedly, $x_{K+2} = F$, $x_{K+3} = F$, and so on. It follows that once the decision maker chooses Arm F, they never return to Arm U. ■

Proposition 4. *An optimal policy involves starting with Arm U.*

Proof. By Propositions 2 and 3, any optimal policy exhibits at most one switch and, once Arm F is chosen, it is chosen in all remaining rounds.

Therefore, to determine the optimal initial action, it suffices to compare two classes of strategies:

- $\sigma_{\bar{F}}$, which prescribes choosing Arm F in every round;
- $\sigma_{U\bar{F}}$, which prescribes choosing Arm U in the first round, continuing with Arm U if a reward is obtained, and switching permanently to Arm F otherwise.

In the first round, both arms return a reward with the same probability, since

$$p\theta_U = \theta_F.$$

If no reward is obtained in the first round (from Arm U under $\sigma_{U\bar{F}}$, from Arm F under $\sigma_{\bar{F}}$), the two strategies $\sigma_{U\bar{F}}$ and $\sigma_{\bar{F}}$ induce identical play from the second round onward.

However, if a reward is obtained in the first round, Proposition 1 implies that $\sigma_{U\bar{F}}$ prescribes choosing Arm U in all remaining rounds, whereas $\sigma_{\bar{F}}$ continues to choose Arm F.

Since Arm U is then known to be good and $\theta_U > \theta_F$, $\sigma_{U\bar{F}}$ strictly dominates $\sigma_{\bar{F}}$ in this contingency.

Because both strategies yield identical outcomes when no reward is obtained and $\sigma_{U\bar{F}}$ yields strictly higher expected utility when a reward is obtained, it follows that

$$EU(\sigma_{U\bar{F}}) > EU(\sigma_{\bar{F}}).$$

Now consider the strategy $U\sigma^*$, which starts with Arm U and follows an optimal continuation strategy thereafter.

By *Proposition 1*, $U\sigma^*$ coincides with $\sigma_{U\bar{F}}$ whenever Arm U returns a reward in the first round, and yields weakly higher expected utility in all other contingencies. Hence,

$$EU(U\sigma^*) \;\geq\; EU(\sigma_{U\bar{F}}) \;>\; EU(\sigma_{\bar{F}}).$$

Therefore, starting with Arm U strictly dominates starting with Arm F. ■

Proposition 5. ***Risk-Neutral Cutoff.*** *A risk-neutral DM explores Arm U for 8 rounds before switching.*

Proof. Consider a utility function $u(w) = w$. By **Proposition 3** (single cutoff) and **4** (starting with Arm U), the optimal switching period K^* satisfies the equation

$$\begin{aligned} p_K\,\theta_U \sum_{t=0}^{N-K} p_U^{\text{good}}(t)\,u(w+tw) \;+\; (1-p_K\,\theta_U) \sum_{t=0}^{N-K} p_F(t)\,u(tw) \\ =\; \theta_F \sum_{t=0}^{N-K} p_F(t)\,u(w+tw) \;+\; (1-\theta_F) \sum_{t=0}^{N-K} p_F(t)\,u(tw), \end{aligned}$$

where $p_X(t)$ is the probability of obtaining t rewards from arm X, and p_K is the Bayesian posterior belief about Arm U being good in round K, after $K-1$ consecutive failures from Arm U. The left-hand side represents the expected utility of exploring Arm U in period K and switching to Arm F if no reward is obtained, while the right-hand side represents the expected utility of switching immediately to Arm F.

For risk-neutral preferences, this simplifies to

$$p_K\,\theta_U\big(w+(N-K)\theta_U w\big) \;+\; (1-p_K\,\theta_U)(N-K)\theta_F w \;=\; (N-K+1)\theta_F w.$$

Solving for the cutoff belief p_K^* gives

$$p_K^* = \frac{\theta_F}{\theta_U\left[1+(N-K)(\theta_U-\theta_F)\right]}.$$

The Bayesian posterior after $K-1$ consecutive failures from Arm U is

$$p_K = \frac{p(1-\theta_U)^{K-1}}{p(1-\theta_U)^{K-1}+(1-p)},$$

where p is the initial prior probability that Arm U is good. The optimal stopping time K^* is the first round where $p_K < p_K^*$.

For the experimental parameters, $\theta_U = 0.2$, $\theta_F = 0.1$ and $N = 30$ (the reward $w = 50$ cancels out), the cutoff equation becomes

$$p_K^* = \frac{0.1}{0.2\,(4 - 0.1K)},$$

and iterating the posterior using Bayes' formula yields $K^* = 9$, since $p_8 = 0.173 > p_8^* = 0.156$ and $p_9 = 0.144 < p_9^* = 0.161$: the DM explores Arm U in rounds 1–8 and switches in round 9. ■

7.9 Three Hypotheses for Over-Switching: Details

7.9.1 Utility Curvature

With utility over earnings, the choice of an expected-utility maximizer in the QU stage depends only on the round, which sets the horizon; total earnings in the problem; and the number of failures observed on Arm U, which sets the belief about it. One problem is paid at random, so earnings in other problems do not enter. Solving the 30-round problem exactly, constant absolute risk aversion never makes a return to Arm U optimal. Constant relative risk aversion can, but only after a reward on Arm F; for relative risk aversion from -20 to 40 and initial wealth up to 1,000 points, no failure on Arm F ever makes a return optimal. For example, with log utility and an initial wealth of 50 points, a subject who switches to Arm F after six failures on Arm U and then receives a reward should return to Arm U. Such returns go in the opposite direction to the response in Table 1, and the gain from them is at most 0.1 to 0.2 points of certainty equivalent, against 50 points per reward. Once Arm U has paid, it pays with probability 0.2 against 0.1 for Arm F, so it first-order stochastically dominates Arm F, and no increasing utility prefers to switch away from it. Rows 3a and 3b of Table 1 hold the round, total earnings and failures on Arm U fixed, with one indicator for each combination; Table 15 reports their logit and Arm U-bad estimates.

Aspiration utilities work differently. With a target, what matters is how far earnings are from it and how many rounds remain. A failure on Arm F leaves earnings unchanged but uses up a round, so a subject who is behind becomes more likely to fall short with Arm F, whose payoff is predictable, and Arm U, which pays twice as often if it is good, becomes more attractive as a gamble. Under a logistic target utility, returns to Arm U are therefore triggered by the lack of rewards as the horizon shrinks: in the exact solutions for the targets and steepness values I fit below, a return can be optimal right after a failure on Arm F (23 of 54 utilities) and never right after a reward. A reward moves the subject toward the target

and makes the gamble less attractive, so this utility predicts less switching after rewards, the sign of the response in Table 1. Under the S-shaped value function, returns can instead follow a reward (62 of 72 utilities), as under constant relative risk aversion. The second argument in the main text therefore does not cover aspiration utilities. The third does, because it holds fixed the rounds left and earnings, the only variables to which a target utility responds; and the simulations below use the fitted aspiration utilities, which include this mechanism.

To check whether curvature can rationalize the choices, I fit standard utilities to each subject's QU-stage decisions (30,491 decisions of 300 subjects). Each decision chooses Arm U with probability $1/(1+\exp\{-\lambda_i[CE_U - CE_F]\})$, where the certainty equivalents of the two arms come from the exact solution of the 30-round problem under the utility and λ_i is a subject-specific precision; parameters are estimated by maximum likelihood on a grid. I fit constant relative risk aversion (CRRA), constant absolute risk aversion (CARA), expo-power utility (Saha 1993; Holt and Laury 2002), which lets relative risk aversion vary with wealth, quadratic utility with satiation, and two aspiration utilities over the problem's earnings: a logistic function of earnings around a target, with the target and its steepness as parameters, and an S-shaped value function around a target, concave above it and convex below it, with its curvature and loss aversion as parameters. Wealth equals the \$3.50 participation fee plus earnings or earnings alone. Curvature fits better than risk neutrality ($p < 0.05$) for 26–50% of the 248 subjects who return to Arm U at least once, against 2–6% of simulated risk-neutral subjects. Under the standard families the fitted parameters are implausible (Table 24, Panel A): more than half of these subjects value an even chance at \$10 or nothing below \$1 or above \$9, and, except under satiation, at most 12% value it between \$2.50 and \$5.50; restricting curvature to plausible values keeps 13–39% of the gain in fit. The aspiration utilities fit better than risk neutrality for 35–42% of these subjects, with more plausible values (25–28% below \$1 or above \$9).

The fitted utilities, however, do not generate the response to rewards. I simulate each of the 147 QU-stage over-switchers ten times over the five problems with their own fitted utility and precision, and estimate rows 2 and 3b of Table 1 on the simulated choices (Panel B). Under every family the estimates are close to zero: between -0.5 and $+1.0$ percentage points per current-wave reward with the controls of row 2, against -3.1 in the data, and between -1.3 and $+2.7$ for two or more rewards, against -10.0; holding the state fixed, between -1.1 and $+1.0$ per reward, against -3.5. The utilities that fit these subjects explain other features of their choices, such as how long they stay on Arm F, but not the response to rewards. The comparison has power: when I simulate 147 subjects whose aspiration utilities do produce a

response of the observed size (−3.0 percentage points per reward and −8.9 for two or more rewards), fitting the same model to their choices and simulating it again recovers most of it (−2.6 and −7.6).

Table 24: Fitted utilities: risk attitudes and the response to rewards

Panel A. Value of an even chance at $10 or nothing (% of the 248 subjects who return)

Utility	Wealth	Below $1	$2.50–$5.50	Above $9	Below $1 or above $9
CRRA	Fee + earnings	36	5	25	61
CRRA	Earnings	32	12	23	55
Expo-power	Fee + earnings	37	5	41	78
Expo-power	Earnings	38	9	30	68
CARA	Either	37	5	35	72
Quadratic, satiation	Either	55	40	0	55
Aspiration, logistic	Earnings	28	27	0	28
Aspiration, S-shaped	Earnings	25	44	0	25

Panel B. Response to current-wave rewards: data and simulated choices (percentage points)

		Controls of row 2		State held fixed (row 3b)	
Utility	Wealth	Per reward	Two or more	Per reward	Two or more
Data		−3.1	−10.0	−3.5	−10.6
CRRA	Fee + earnings	+0.1	+0.8	+0.5	+1.5
CRRA	Earnings	−0.3	−1.3	+0.5	+1.0
Expo-power	Fee + earnings	0.0	−0.4	+0.2	−0.1
Expo-power	Earnings	+0.2	+0.3	+1.0	+1.9
CARA	Either	−0.1	0.0	+0.1	+0.3
Quadratic, satiation	Either	+1.0	+2.7	−1.1	−2.3
Aspiration, logistic	Earnings	−0.5	−0.4	−0.7	−1.2
Aspiration, S-shaped	Earnings	+0.1	+0.4	−0.2	+0.3

Notes: Maximum-likelihood fits to each subject's QU-stage decisions in Treatment 1 (30,491 decisions, 300 subjects); percentages over the 248 subjects who return to Arm U at least once. Each decision chooses Arm U with a logistic probability in the difference between the certainty equivalents of the two arms, from the exact solution of the 30-round problem, with a subject-specific precision. Wealth: the $3.50 participation fee plus earnings in the problem, or earnings alone; CARA does not depend on wealth. A risk-neutral subject values the lottery at $5. About a third of the CRRA fits lie at the edge of the grid (relative risk aversion from −20 to 40 with the fee, −20 to 0.99 without); all of them are already in the first or third column. Quadratic utility with satiation cannot be risk seeking. Aspiration utilities are fitted to earnings alone; a target relative to the fee gives the same fit. Panel B: the first row is the data; in the other rows, each of the 147 QU-stage over-switchers is simulated ten times over the five problems with their own fitted utility and precision, and rows 2 and 3b of Table 1 (and two or more rewards relative to none, with the same controls) are estimated on the simulated Arm F decisions in the QU stage; standard errors of the simulated estimates are 0.3–0.4 per reward and 0.8–1.2 for two or more rewards.

7.9.2 Execution and Valuation Distortions: Formal Framework and Propositions

Formal model. The canonical model has the agent solve a single maximization problem each period: choose the arm with the higher value $V_t(a)$, $a \in \{U, F\}$, obtained by weighing the immediate payoff against what continuing to explore is worth going forward. I introduce two distortions on top of this problem. A wave ω is defined as the sequence of consecutive draws from the same arm, so it only ends at the moment a different arm is chosen; and h_t^ω is the local history within that wave, as opposed to h_t, the problem's full history.

The first distortion, which I call the *execution distortion*, is *additive*: it is an extra term added to the agent's valuation $V_t(a)$, which leaves the canonical valuation and $m(h_t)$ unchanged and enters V_t as the current d pull, held constant on the current arm; it only distorts the agent's assessment of the current arm relative to the other, from the perspective of the current period. This distortion has two components. The first, $d(a, h_t^\omega)$, the *d pull*, captures the distortion generated by what has happened so far *within the current wave* — the draws so far, the rounds spent in that wave. The agent judges the d pull from the perspective of the current period: every round of the current arm that they consider carries the current d pull. The second, $D(a, S, h_t^\omega)$, captures the demand for flexibility: the agent anticipates that within-wave experience could make them change their mind *later in the wave*, deviating from what V_t prescribes, and so values the flexibility to react to it; committing to an arm for the rest of the problem removes that flexibility. I therefore impose $D(a, 1, h_t^\omega) = 0$, $D(a, S, h_t^\omega) \leq 0$, and $|D(a, S, h_t^\omega)|$ non-decreasing in S: a one-round choice keeps full flexibility, and each additional round committed can only limit flexibility. Here S is the horizon of the current choice, the number of rounds ahead it covers. Whether this within-wave distortion enters the agent's decision at all, and whether a demand for flexibility enters along with it, is indexed by two parameters, η and $\beta \in \{0, 1\}$: η turns the whole within-wave channel on or off, and β, active only when $\eta = 1$, governs whether D enters. I say that the execution distortion comes with a *demand for flexibility* when $\beta = 1$.

The second distortion, the *valuation distortion* $m(h_t)$, works differently: rather than being added to the agent's valuation, it enters *inside* $V_t(a)$ itself, directly changing how the agent values exploring or exploiting the arms. It is stable across waves and persists when the agent switches arms; it may reflect, for example, an inherent preference for switching or an incorrect idea of how to maximize expected payoffs. Whether this distortion is active is indexed by a third parameter, $\gamma \in \{0, 1\}$. I define $m(h_t)$ as a distortion under which the agent's switching rule, the contingent rule that maximizes $V_t\big(a, \gamma m(h_t)\big)$, is not single-

switching before or after Arm U pays; a distortion that leaves the rule single-switching is observationally equivalent to $\gamma = 0$. I say that it affects the QU stage in the first case and the QC stage in the second; since $m(h_t)$ depends on the full history, it can affect either stage or both, and it affects at least one. I assume that the distorted problem is time-consistent: the rule that is best at the start of the problem remains best, judged with $V_t\big(a, \gamma m(h_t)\big)$, at every later history, so the rule an agent would pick before the problem starts is the rule that guides their later choices. The valuation distortion can coexist with either form of the execution distortion.

Putting the pieces together, the agent's period-t choice solves

$$(a_t, S_t) = \arg \max_{a \in \{U,F\},\, S} V_t\big(a, S, \gamma\, m(h_t)\big) \;+\; \eta\, d(a, h_t^\omega) \;+\; \eta\, \beta\, D(a, S, h_t^\omega),$$

with η, β, γ governing which distortions come into play, where x_t is the arm of the current wave (for $a \neq x_t$, both D and d are evaluated at the empty history $h_0^{\omega'}$ of the wave that choosing a would start, so $d(a, h_0^{\omega'}) = 0$ by (P1) below), and where $V_t(a, S, \cdot)$ is the value of drawing arm a for the next S rounds and following the best policy thereafter, as evaluated by the agent from the perspective of the current period, with the same constant $\eta\, d(x_t, h_t^\omega)$, independent of S, added to the value of arm x_t in every future period. With $\gamma = 0$ and $d = 0$, this is the canonical model's value, which I write $V_t(a, S)$. The round-by-round choice case is included as $S = 1$, and I write $V_t(a, \cdot)$. Finally, when several block lengths yield the same value, the agent chooses the longest of them; this matters only when blocks can last more than one round.

I impose four properties on the structure of $d(a, h_t^\omega)$ and $D(a, S, h_t^\omega)$. First (P1), $d(a, h_0^\omega) = C$: at the beginning of each wave, when no wave-history has been accumulated, the d pull is constant, which I normalize WLOG at $C = 0$. Second (P2), if $\omega \neq \omega'$, then $d(a, h_t^{\omega'} \mid h^\omega) = d(a, h_t^{\omega'})$ and $D(a, S, h_t^{\omega'} \mid h^\omega) = D(a, S, h_t^{\omega'})$: distortion is fully local, and no component is carried across waves. Third (P3), the first Arm U reward resets the d pull: if ω is an Arm U wave and the last draw in h_t^ω is the first reward from Arm U, then $d(U, h_t^\omega) = 0$. This aligns the d pull with the standard part of the model, in which a breakthrough reward significantly impacts the agent's valuation of Arm U. Otherwise, the sign of d is unrestricted: for the arm a of the current wave ω, within-wave experience may push the agent toward leaving it, $d(a, h_t^\omega) < 0$, as with frustration or boredom, or toward staying, $d(a, h_t^\omega) > 0$, as with inertia, and random errors may push in either direction; choosing the other arm, a', would begin a new wave ω', so its term is $d(a', h_0^{\omega'}) = 0$. Fourth (P4), d and D depend only on outcomes the agent experiences or expects to experience: when

the agent neither experiences nor expects to experience any within-wave outcome, as under the Non-Experienced Choice structure below, both are zero. These properties govern the d pull the agent actually experiences: once the agent leaves an arm, its d pull is gone (P2), even though at the moment of choice they judged the arm by it, and the first Arm U reward resets it (P3).

Since the execution distortion responds only to within-wave experience, a *choice architecture* (a protocol that changes when and how often the agent makes decisions, and what they have observed at each decision point) that limits or removes this experience at the switching decision limits or shuts it off. I formalize two, which do so to different degrees.

Definition: Block Choice. Under Block Choice, whenever the agent chooses an arm, they also choose a number of periods to automatically draw from that arm, forgoing the opportunity to revise the arm during that block. Thus, the agent chooses sequences of same-arm draws rather than an arm in every round. If Arm U returns its first reward, the block restarts as the informational stage changes. I implement this architecture in Treatment 4 (Block Choice).

Proposition 6. *If an agent is only distorted by within-wave experiences ($\gamma = 0$ and $\eta = 1$) and the agent has no demand for flexibility ($\beta = 0$), then under a Block Choice structure the agent follows a single-switching path, exploring Arm U for at least as long as the baseline model prescribes unless Arm U pays earlier. If, in addition, $d(U, h_t^\omega) \leq 0$ whenever ω is an Arm U wave containing only failures, the agent follows the optimal switching path.*

Proof. Recall that $V_t(a, S)$ is defined as the value under the baseline model's optimal policy — the value obtained from solving the baseline dynamic program by backward induction over the remaining horizon, absent any distortion (it already encodes the solution to that optimization at every t and every history h_t).

With $\gamma = 0$, the valuation distortion term is inactive, and with $\beta = 0$ the period-t objective reduces to

$$V_t(a, S, 0) + d(a, h_t^\omega).$$

At the start of any wave ω, by construction $h_t^\omega = h_0^\omega = 0$, and by the normalization in (P1), $d(a, h_0^\omega) = C = 0$ for the arm a of that wave, so the d pull vanishes. At $t = 0$ there is no current wave and $d = 0$, and the period-t objective reduces exactly to

$$V_t(a, S),$$

for which the agent chooses Arm U and a block length S. Since a block of S rounds is just a

sequence of consecutive Arm U choices, choosing $a = U$ and the optimal S at the beginning of the first wave is identical to solving the canonical problem through backward induction, because the objective function is also $V_t(a, S)$.

At the outset of the problem ($t = 0$, the start of wave 1), $V_0(a, S)$ is the baseline model's value, whose solution — established in the baseline model — is to choose Arm U and continue choosing it until either a reward is realized, or accumulated unfavorable evidence reaches the point at which continued exploration is no longer optimal by backward induction; call this stopping point U^*. Every block length up to U^* yields the same $V_0(U, S)$, since the baseline policy keeps drawing Arm U until U^* anyway, while any longer block lowers it; since the agent chooses the longest block when indifferent, they pick $S = U^*$. The agent's wave-1 commitment therefore reproduces U^* exactly, as in the baseline model.

Suppose no reward has arrived by U^*. When the block ends, the agent makes a new block decision while still in the Arm U wave ω, so they evaluate every option with each round of Arm U shifted by the current d pull $d(U, h_t^\omega)$, at the history h_t inherited from wave ω (specifically, that U^* rounds have elapsed with no reward). By Bayesian updating over the entire problem, V_t prescribes switching to Arm F after U^* unsuccessful rounds. If $d(U, h_t^\omega) \leq 0$, it can only make Arm U less attractive, and the agent switches to Arm F. If $d(U, h_t^\omega) > 0$, the agent may extend exploration with further Arm U blocks for as long as the d pull outweighs the gain from switching, until Arm U returns its first reward (the reward case below) or the problem ends. In every case, exploration lasts at least U^* rounds. When the agent switches to Arm F, V_t favors Arm F even with the current d pull $d(U, h_t^\omega)$ added to Arm U, and Arm F carries no d pull of its own, since $d(F, h_0^{\omega'}) = 0$. Looking ahead, nothing brings the agent back: Arm F's draws reveal nothing about Arm U, and the d pull on Arm U is held constant, so returning to Arm U is no more attractive later than it is now. Evaluated at the moment of the switch, the best continuation therefore stays on Arm F for the rest of the problem, and since the agent chooses the longest block when indifferent, they choose a single block to the end, rather than a shorter block with the intention of repeating it later, so no further decision arises.

Hence, absent a reward, the agent's realized path is: Arm U for at least U^* periods, followed by Arm F until the problem ends — a single-switching path, which coincides with the baseline model's optimal switching path when $d(U, h_t^\omega) \leq 0$.

Suppose instead Arm U returns its first reward, whether during the first block or during a later block. Per the Block Choice structure, a new block must then be chosen. By (P1), the Arm F wave term is $d(F, h_0^{\omega'}) = 0$, and by (P3), the Arm U wave term was reset to

$d(U, h_t^\omega) = 0$, so the decision is driven by $V_t(U, 1)$ versus $V_t(F, 1)$ alone, which favor Arm U for the rest of the problem by first-order stochastic dominance and monotonicity. Since the agent chooses the longest block when indifferent, they choose a single block to the end. Since later rewards do not restart the block, no further decision arises, and the agent remains on Arm U for the rest of the problem.

In both cases, the agent switches at most once and explores Arm U for at least U^* rounds unless it pays earlier; when $d(U, h_t^\omega) \leq 0$ on Arm U waves containing only failures, the realized path coincides exactly with the baseline model's optimal switching path. ■

Remark 1. If $\beta = 1$, the agent's wave-start choice no longer reduces to $V_t(a, S)$ alone: the flexibility term $D(a, S, h_0^\omega)$, which need not vanish at a fresh wave even though d itself does, re-enters the objective, so an agent with a demand for flexibility may rationally choose a block length shorter than U^* if they anticipate that continuing exploration could generate distortion severe enough later in the wave that they would rather stop. Unlike in models of present bias, where an agent who anticipates future deviations values commitment, the d pull here reflects boredom or frustration that the agent actually bears, so such an agent values the flexibility to react to it, which a shorter block provides. This is the only way an agent without a valuation distortion can over-switch under Block Choice. If $\gamma = 1$, the distortion is in V_t itself, so the agent chooses blocks that reflect it, instead of the block structure shutting it down.

Remark 2. Because D is zero for a one-round block and never positive, the demand for flexibility does not change the arm chosen where the d pull of both arms is zero, at the start of the problem and right after Arm U first pays: there, a one-round block on the arm V_t prescribes is strictly better than any block on the other arm.

Definition: Non-Experienced Choice. A choice architecture is called non-experienced if the agent commits, ex ante and in a strategy-method format, to a complete contingent rule governing play for the entire problem, and observes no within-wave outcomes at any point before the problem ends — only the final realized payoff is revealed once play concludes.

Proposition 7. *If $\gamma = 0$, an agent follows the optimal single-switching path under a Non-Experienced Choice structure, regardless of η and β.*

Proof. Under Non-Experienced Choice the agent chooses the rule once, ex ante, and observes no within-wave outcomes, so $d = D = 0$ by (P4), whatever η and β.

With $\gamma = 0$, the valuation distortion term is likewise inactive, so the agent's problem

reduces to choosing, once and for all, the contingent rule that maximizes

$$V_t(a)$$

at every t and every history h_t the rule might encounter — exactly the baseline model's problem, whose solution is the optimal single-switching path. ■

Propositions 6 and 7 characterize how an agent with only execution distortions reacts to the different choice architectures, but they are silent about agents with valuation distortions. Proposition 8 covers them.

Proposition 8. *Consider an agent who over-switches in a stage of the free-choice problem.*

(a) An over-switcher becomes a single-switcher in their respective stage under Non-Experienced Choice if and only if their over-switching in that stage comes only from execution distortions.

(b) A subject's switching structure (single or multiple waves on each arm) is consistent across the two architectures, up to the differences that the execution distortion can create at the decisions where it is allowed to act.[49]

Statement (b) holds for any η and β.

Proof. *Part (a).* By the proof of Proposition 7, the execution distortion vanishes under Non-Experienced Choice, so the agent follows their *switching rule*: the contingent rule that maximizes $V_t\big(a, \gamma m(h_t)\big)$ at every history, staying on the current arm when indifferent and starting with Arm U. This rule is single-switching in a stage if and only if the valuation distortion does not affect that stage; in that case, if $\eta = 0$, the agent would follow the rule in free choice too and not over-switch, so their over-switching comes only from execution distortions.

Part (b). Under Non-Experienced Choice the subject follows their rule (Part (a)). Under Block Choice, at the decisions made without within-wave experience, the start of the problem and right after Arm U first pays, the d pull is zero and, by Remark 2 after Proposition 6, the subject chooses the arm their rule prescribes. Differences from the rule can therefore arise only at decisions where the d pull acts. ■

Corollary 1. After Arm U pays, under Block Choice a subject chooses the arm their rule prescribes, and chooses a single Arm U block to the end of the problem only if, from

[49] I set aside the starting arm, which Ex Ante Rules do not let me observe, since every rule in the menu starts with Arm U.

that point, their rule stays on Arm U to the end of the problem whatever the draws. This holds for any η and β.

Proof. After Arm U first pays, the block restarts. At that decision the d pull is zero for both arms, by (P1) and (P3), so the subject evaluates the options with $V_t\big(a, S, \gamma m(h_t)\big) + \eta\beta D(a, S, h_t^{\omega})$, and by Remark 2 after Proposition 6 the demand for flexibility does not change the arm: the subject chooses the arm their rule prescribes. A single Arm U block to the end of the problem is chosen only if no shorter Arm U block is better; since D can only penalize longer blocks, this requires $V_t\big(\cdot, \gamma m(h_t)\big)$ to keep the subject on Arm U for the rest of the problem whatever the draws, that is, from that point, their rule stays on Arm U to the end of the problem whatever the draws. The argument uses only the decision right after the first reward, where the d pull is zero, so it requires neither $\eta = 0$ nor $\beta = 0$. ■

7.9.3 Block Choice, Additional Details

Problems in which Arm U never paid.

- Transition matrix, switching category (T1 → Block), Treatment 1 over-switchers, $n = 32$:

	Block: under	Block: single	Block: over	Total
T1: over-switcher	3	16	13	32

Problems in which Arm U paid.

- Transition matrix, QC-stage switching category (T1 → Block), Treatment 1 QC-stage over-switchers, $n = 21$; "left at once": left Arm U right after it paid and did not return:

	Block: left at once	Block: 0 (single)	Block: over	Total
T1: over-switch	2	13	6	21

First Arm U wave, matched sample. Restricted to subjects classified as over-switchers in Treatment 1 and observed in both treatments ($n = 41$); one further subject chose only Arm F in their first Block Choice problem and has no Arm U wave. Under Block Choice, the first Arm U wave is measured by the Arm U block lengths the subject chose (a single block for 34 of the 41 subjects), so it is observed even when Arm U pays during the wave. Its mean is 3.76 rounds in Treatment 1 vs. 10.02 rounds under Block Choice

($t = -5.39$, $p < 0.001$). This holds for subjects starting with either arm in Treatment 1, though the shift is largest for subjects who started with Arm F ($n = 10$): 2.70 vs. 14.20 rounds ($t = -3.73$, $p = 0.005$), compared to those who started correctly with Arm U ($n = 31$): 4.10 vs. 8.68 rounds ($t = -4.42$, $p < 0.001$). Eleven of these 41 changed their starting arm between treatments, 8 of them from Arm F to Arm U.

Block lengths. Among subjects who become single-switchers under Block Choice, 83% (10 of 12) of those who switch to Arm F in the QU stage do so in a single Arm F block, and 77% (10 of 13) stay on Arm U in a single block after it pays in the QC stage. For Arm U exploration in the QU stage, the share choosing a single block is 50% (7 of 14).

Combining the two treatments: details. Throughout, I compare Block Choice behavior with what the rule chosen under Ex Ante Rules does, not with its exact thresholds or lengths: rules were chosen after Block Choice, from a restricted menu, so their parameters are only approximate.[50] Subjects with a single-switching ex ante rule were more likely to single-switch under Block Choice: 10 of 17, against 5 of 14 of those with another rule, in the QU stage, and 10 of 12 against 3 of 9 in the QC stage (one-sided Fisher exact $p = 0.18$ and $p = 0.03$).

QU stage. Two groups of subjects with a single-switching ex ante rule are consistent with a demand for flexibility (Remark 1 after Proposition 6). Five split their exploration into several blocks but still single-switch: they keep the option to change course, but keep choosing the same arm after each block until their single switch. Five others over-switch under Block Choice: they split their exploration into several blocks and, at the end of a block, behave as if responding to an execution distortion.

Five subjects with another ex ante rule nonetheless single-switch under Block Choice, which the framework can accommodate. Four switch at a decision made with within-wave experience, so they behave as if an execution distortion brought them to a single switch: two chose a 15-round cycle, one chose adaptive thresholds and switches after several blocks, and one chose a consecutive-failures rule and plays one block on each arm. The fifth chose a 30-round cycle and plays a single Arm U block for the whole problem, as if following their cycle, which never leaves Arm U within the problem.

Two subjects show behavior the framework cannot accommodate. One, with a single-switching rule, plays a single 30-round Arm F block, committing at the start of the problem, without within-wave experience, never to explore Arm U. The other, with another rule, never

[50]The one exception is the 30-round cycle, chosen by a single subject in both stages: it lasts the whole problem, so within the problem it never leaves Arm U, and I compare their behavior with that.

chooses Arm U, playing Arm F blocks of several rounds from the start of the problem.[51]

QC stage. Subjects whose rule locks in Arm U are consistent with Corollary 1 when they choose Arm U right after it pays. Ten stay on Arm U to the end, three of them in several blocks, consistent with a demand for flexibility that they end up not using; one stays on Arm U in a block shorter than the rest of the problem and leaves it later, as if responding to an execution distortion at the end of that block. Among subjects with another rule, two behave as if following their rule within the problem: one chose a 30-round cycle and stays on Arm U to the end, and one leaves Arm U as soon as it pays, two rounds before the end, as their rule, which keeps cycling after Arm U pays, also does.

The framework cannot accommodate three subjects, all contradicting Corollary 1: one leaves Arm U as soon as it pays although their rule locks it in, and two commit to Arm U to the end although their rule leaves it.

In total, five subject-stages are not consistent with the framework: the two QU-stage and three QC-stage subjects described above.

7.9.4 Robustness Checks for Block Choice and Ex Ante Rules Results

Two concerns arise because subjects in Block Choice and Ex Ante Rules are selected on over-switching in one Treatment 1 problem, and both treatments are always played after Treatment 1. First, regression to the mean: a subject labeled an over-switcher on the basis of a single problem may simply revert to their usual behavior. Second, learning: the improvement may continue a trend already present in Treatment 1.

To address both, I split Treatment 1 over-switchers into *fixers* and *keepers*: in the terms of the main text, subjects who become single-switchers under the treatment and subjects who continue to over-switch under it. If fixers are a product of regression to the mean, they should rarely over-switch in their other Treatment 1 problems; if they are learning, their behavior should improve across Treatment 1 problems. Tables 25–28 report, for each group, medians in each Treatment 1 problem and under the treatment (Panel A), and the share of the group's other Treatment 1 problems with over-switching (Panel B).

Block Choice. In the QU stage, fixers over-switch in 81% of their other Treatment 1 problems (keepers 98%), with a median of 2–4 extra switches and a first Arm U wave of 2–6 rounds, against 3–10 extra switches and 1–3 rounds for keepers, and no trend in the last

[51] The other subject with a single-switching rule who never chooses Arm U plays thirty one-round Arm F blocks, so the execution distortion can keep them on Arm F at every decision after the first; I do not test the starting arm.

three to four of the five problems. In the QC stage, fixers over-switch in 89% of their other problems (keepers 82%) and leave Arm U 3–6 rounds after it pays.

Ex Ante Rules. In the QU stage, fixers over-switch in 80% of their other Treatment 1 problems (keepers 100%), with a median of 1–5 extra switches and a first Arm U wave of 2–6 rounds, against 2.5–7 extra switches and 2–3 rounds for keepers, and no trend in the last three to four of the five problems. In the QC stage, they over-switch in 79% (keepers 89%) and leave Arm U 1–9 rounds after it pays.

In both treatments and stages, fixers over-switched persistently in Treatment 1, which is inconsistent with regression to the mean, and their extra switches show no trend in the last three to four of the five problems. In the QU stage, this is also inconsistent with learning. In the QC stage, fixers stay on Arm U longer after it pays in the later Treatment 1 problems, although each median rests on 3 to 8 subjects, so I cannot rule out some learning in that stage.

Table 25: Block Choice, QU stage: behavior by group

	Fixers			Keepers		
	n	Extra sw.	First U	n	Extra sw.	First U
Panel A: Medians						
T1, Problem 1	13	4	2	12	7	1.5
T1, Problem 2	11	4	5	11	3	3
T1, Problem 3	11	2	6	10	4	2.5
T1, Problem 4	11	2	5	9	10	1
T1, Problem 5	13	2	5	11	6	3
Block Choice	16	0	10	13	2	7
Panel B: Other Treatment 1 problems						
Share over-switching		0.81 (43)			0.98 (40)	

Notes: Groups are defined by switching in the last Treatment 1 problem in which Arm U never paid and the last Block Choice problem in which Arm U never paid. Extra sw.: switches beyond the single-switching prediction. First U: length of the first Arm U wave. Treatment 1 rows include problems in which Arm U never paid, and problems in which Arm U paid but the first Arm U wave ended before it paid or lasted at least eight rounds, the risk-neutral benchmark. Panel B: share of the group's other Treatment 1 problems with over-switching; number of problems in parentheses.

7.10 Classification of Treatment 1 Over-Switchers

This appendix gives the details behind the classification of Treatment 1 over-switchers into noisy-type and valuation-type groups.

Table 26: Block Choice, QC stage: behavior by group

	Fixers			Keepers		
	n	Extra sw.	Stay U	n	Extra sw.	Stay U
Panel A: Medians						
T1, Problem 1	7	6	3	5	3	2
T1, Problem 2	6	3	3	4	2.5	3.5
T1, Problem 3	8	5	3.5	3	3	1
T1, Problem 4	4	2.5	6	3	10	1
T1, Problem 5	6	4.5	6	2	16	1.5
Block Choice	13	0	24	6	1	7
Panel B: Other Treatment 1 problems						
Share over-switching		0.89 (18)			0.82 (11)	

Notes: Groups are defined by switching after Arm U pays in the last Treatment 1 problem in which Arm U paid and the last Block Choice problem in which Arm U paid. Extra sw.: switches after Arm U pays beyond the single-switching prediction. Stay U: rounds on Arm U after it pays before first leaving it. Treatment 1 rows include all problems in which Arm U paid. Panel B: share of the group's other Treatment 1 problems in which Arm U paid that show over-switching; number of problems in parentheses.

Table 27: Ex Ante Rules, QU stage: behavior by group

	Fixers			Keepers		
	n	Extra sw.	First U	n	Extra sw.	First U
Panel A: Medians						
T1, Problem 1	13	5	2	14	7	2
T1, Problem 2	13	2	5	11	6	3
T1, Problem 3	11	2	6	12	2.5	2.5
T1, Problem 4	11	1	2	10	5.5	3
T1, Problem 5	14	2	4.5	11	4	3
Ex Ante Rules	17	0	9	14	1	10
Panel B: Other Treatment 1 problems						
Share over-switching		0.80 (45)			1.00 (44)	

Notes: Groups are defined by switching in the last Treatment 1 problem in which Arm U never paid and the rule chosen under Ex Ante Rules, among subjects with a Block Choice problem in which Arm U never paid. Variables as in Table 25. In the Ex Ante Rules row, Extra sw. equals 1 if the chosen rule is not single-switching and 0 otherwise, and First U is the first Arm U wave the rule prescribes.

Table 28: Ex Ante Rules, QC stage: behavior by group

	Fixers			Keepers		
	n	Extra sw.	Stay U	n	Extra sw.	Stay U
Panel A: Medians						
T1, Problem 1	7	6	2	6	4.5	3
T1, Problem 2	5	2	1	6	3	3.5
T1, Problem 3	7	1	5	5	7	1
T1, Problem 4	3	1	9	6	10.5	1
T1, Problem 5	4	4.5	6	4	9	3
Ex Ante Rules	12	0	–	9	1	–
Panel B: Other Treatment 1 problems						
Share over-switching		0.79 (14)			0.89 (18)	

Notes: Groups are defined by switching after Arm U pays in the last Treatment 1 problem in which Arm U paid and the QC rule chosen under Ex Ante Rules, among subjects with a Block Choice problem in which Arm U paid. Variables as in Table 26. In the Ex Ante Rules row, Extra sw. equals 1 if the chosen rule for after Arm U pays is not single-switching and 0 otherwise; – not defined for a rule.

Anchors and the other treatment responses. Table 29 cross-tabulates QU-stage behavior under Block Choice and Ex Ante Rules for the 31 Treatment 1 over-switchers who played both. Classifying the 13 who are not anchors from their Treatment 1 behavior, like every other over-switcher, yields 6 noisy-type and 7 valuation-type subjects. In the QC stage, Panel B of Table 2 gives the same cross-tabulation for the 21 Treatment 1 QC-stage over-switchers who played both treatments: 10 stay on Arm U under both and 5 keep switching under both (the QC-stage anchors), 4 respond differently to the two treatments, and 2 switch away from Arm U as soon as it pays under Block Choice. Since the QC-stage anchors cannot be told apart in free switching, these 6 are not classified.

Table 29: Treatment responses and Treatment 1 classification

Block Choice	Ex Ante Rules	n	Noisy type	Valuation type
Single-switching	Single-switching rule	10 (anchors)	10	0
Over-switching	Over-switching rule	8 (anchors)	2	6
Single-switching	Over-switching rule	5	3	2
Over-switching	Single-switching rule	5	2	3
Never chooses Arm U	Either	3	1	2

Notes: Treatment 1 over-switchers who played Treatments 4 and 5, QU stage. Columns 4–5: group assigned by the nearest-centroid rule on Treatment 1 behavior.

Choice of markers. Table 30 compares the two anchor groups on 21 candidate measures of their Treatment 1 behavior. Failure-reactivity refers to the share of failures immediately followed by a switch. The two groups differ at the 5% level on 10 of the measures with 20,000 permutations per measure, and on 9 with the 2,000 permutations used for the random groups below. The differences run in one direction: valuation-distortion anchors switch more, in shorter waves, and react more to failures; the two groups do not differ in comprehension errors.

All the measures on which the anchors differ. With the 9 measures that differ at the 5% level with 2,000 permutations as markers, standardized and equally weighted, 13 of the 18 anchors are assigned to their own group (13 when each is left out), and the label of 12 of the 147 subjects changes (90 noisy type and 57 valuation type). A plan is then the best model for 62% of noisy-type and 38% of valuation-type subjects ($p = 0.004$); the accuracy of the best plan model against the best model without a plan is 65.9% vs. 60.1% for noisy-type and 51.7% vs. 52.3% for valuation-type subjects; and the advantage of the plan is larger for noisy-type subjects ($p < 0.001$).

Table 30: Treatment 1 behavior of the anchors

Measure	Noisy execution (10)	Valuation distortion (8)	Permutation p	Cliff's δ
Switch rate (marker)	0.18	0.35	0.026	−0.63
Number of switches	5.4	10.5	0.028	−0.63
Number of waves	6.4	11.5	0.028	−0.63
Extra switches beyond single switching	4.0	9.5	0.020	−0.65
Share of problems with optimal switching (marker)	0.50	0.18	< 0.001	+0.85
Number of Arm U waves	3.1	6.0	0.019	−0.69
Number of Arm F waves	3.3	5.5	0.056	−0.50
Mean Arm U wave length	4.1	2.6	0.090	+0.49
Median Arm U wave length	3.9	2.6	0.233	+0.34
Mean Arm F wave length	7.0	4.1	0.055	+0.55
Median Arm F wave length	6.9	3.4	0.0495	+0.53
Share of Arm F waves lasting one round	0.13	0.35	0.156	−0.36
Length of the first wave	5.1	2.4	0.054	+0.51
Length of the first Arm U wave	5.6	2.4	0.046	+0.50
Length of the last Arm F wave	7.8	6.9	0.747	0.00
Failure-reactivity, average of the two arms (marker)	0.21	0.40	0.028	−0.53
Failure-reactivity on Arm U	0.27	0.50	0.041	−0.45
Failure-reactivity on Arm F	0.15	0.31	0.069	−0.63
Starts on Arm U (share)	0.60	1.00	0.093	−0.40
Ends on Arm U (share)	0.20	0.50	0.326	−0.30
Comprehension-quiz errors	2.6	2.9	0.900	−0.21

Notes: Means. All measures except the share of problems with optimal switching and comprehension errors are computed in the subject's last Treatment 1 problem in which Arm U never paid; lengths in rounds. A wave is a sequence of consecutive choices of the same arm. Permutation p: two-sided test of equal means, 20,000 permutations. Cliff's δ compares noisy-execution with valuation-distortion anchors (positive: larger for noisy execution). Bold: the three classification markers.

Comparison with random groups. I draw 10,000 random splits of the 147 Treatment 1 over-switchers into groups of 10 and 8 and apply exactly the same tests to each: only 3.4% differ on at least 9 of the 21 measures, and 90% differ on fewer than three. For each draw I then rerun the whole procedure: I keep the measures that pass, build every combination of two or three of them, classify the 147 subjects with each combination, and keep the best one. Draws in which fewer than three measures pass count as doing worse than the anchors; these are 9,022 of the 10,000 draws (90%). Draws that differ on as many measures as the anchors (338, or 3.4%) mostly recover the same groups: the median one puts about 125 of the 147 subjects in the same group as my classification. Among the 978 draws (9.8%) in which at least three measures pass, the p-values for the share of anchors assigned to their own group, the labels reproduced in a held-out problem and the labels recovered by consensus clustering are 0.073, 0.39 and 0.60, respectively. Counting all 10,000 draws, the p-values are 0.007, 0.038 and 0.059.

Predicting the anchors' response. The random relabelings in the main text split the 18 anchors into groups of 10 and 8; across 20,000 of them, 8.9 anchors are placed in their own group on average. As a stricter check, each time an anchor is left out I choose the three measures of Table 30 that best separate the other 17 (largest absolute Welch t-statistics) and predict the left-out anchor from them. This places 14 of the 18 in their own group; 4.9% of 5,000 relabelings, with the measures chosen anew in each, do as well.

QC-stage anchors. Table 31 compares the 10 and 5 QC-stage anchors on 19 measures of their Treatment 1 behavior, computed, as for the QU-stage anchors, in the problem used to classify them: the QC stage of their last problem in which Arm U paid. None differs at the 5% level. With 10 and 5 subjects, a two-sided t-test at the 5% level detects a difference of 1.7 standard deviations in a single measure with probability 0.8; the two groups differ by 0.53, 0.01 and 0.36 standard deviations in switch rate, failure-reactivity and the share of optimal problems (Cohen's d).

Stability with respect to the anchors. Sixteen of the 18 anchors are assigned to their own group; the other two are valuation-distortion anchors assigned to the noisy-type group. (i) Leaving out one anchor at a time and recomputing the centroids changes the classification of the 147 subjects by 2.1% on average (at most 6.8%). (ii) Resampling the two anchor groups 1,000 times, out-of-bag anchors are misassigned 15.1% of the time. The classification of the 147 subjects differs from my classification by 7.1% on average (99th percentile 18.4%), but

Table 31: Treatment 1 behavior of the QC-stage anchors

Measure	Noisy execution (10)	Valuation distortion (5)	Permutation p	Cliff's δ
Switch rate	0.22	0.34	0.404	+0.04
Number of switches	4.8	8.6	0.299	−0.12
Extra switches beyond single switching	4.8	8.4	0.330	−0.12
Share of problems in which Arm U paid that show optimal QC-stage switching	0.05	0.11	0.500	−0.26
Number of Arm U waves	3.2	4.8	0.416	−0.04
Number of Arm F waves	2.6	4.8	0.253	−0.18
Mean Arm U wave length	5.2	5.0	0.927	+0.22
Median Arm U wave length	4.8	5.0	0.954	+0.18
Mean Arm F wave length	3.4	3.6	0.975	−0.42
Median Arm F wave length	3.3	3.7	0.944	−0.46
Share of Arm F waves lasting one round	0.52	0.20	0.207	+0.42
Length of the first Arm U wave	5.4	5.2	0.962	0.00
Length of the first Arm U wave, share of the QC stage	0.24	0.20	0.730	0.00
Length of the last Arm F wave	3.4	3.4	1.000	−0.30
Failure-reactivity, average of the two arms	0.36	0.37	0.985	+0.20
Failure-reactivity on Arm U	0.22	0.42	0.153	−0.20
Failure-reactivity on Arm F	0.51	0.31	0.373	+0.24
Ends on Arm U (share)	0.60	0.20	0.283	+0.40
Comprehension-quiz errors	2.7	2.8	1.000	+0.04

Notes: Means. QC-stage anchors: Treatment 1 QC-stage over-switchers who stay on Arm U under both Block Choice and Ex Ante Rules (noisy execution) or keep switching under both (valuation distortion). All measures except the share of problems in which Arm U paid that show optimal switching, and comprehension errors, are computed in the QC stage of the subject's last Treatment 1 problem in which Arm U paid; lengths in rounds. Permutation p: two-sided test of equal means, 20,000 permutations. Cliff's δ compares the noisy-execution with the valuation-distortion group (positive: larger for noisy execution).

the majority label across resamples matches my label for every subject; 10.9% of subjects (16/147) disagree with it in at least a quarter of the resamples, and 21.8% (32/147) in at least a tenth.

Stability of the classification. The checks below show that the label is a stable feature of the subject: it holds across problems, across classification methods and after the restricted treatments. The interpretation of the two groups rests on the anchors, whose labels come from the treatments.

- *Reliability.* Across the 392 problems of the 147 over-switchers in which Arm U never paid, the intraclass correlation of a single problem's value is 0.67 for switching rate and 0.64 for failure-reactivity: about two thirds of the variation is between subjects. Across all Treatment 1 problems, including the QU stage of problems in which Arm U paid, they are 0.52 and 0.49.

- *Unsupervised recovery.* Across 300 random draws that vary both the feature subset (drawn from a pool of 16 behavioral measures) and the clustering algorithm, subjects with the same label co-cluster in 63–72% of draws, against 44% for subjects with

different labels. The consensus partition recovers the classification for 87.8% of subjects ($\kappa = 0.74$).

- *Held-out problem.* In subjects' last Treatment 1 problem in which Arm U paid, the two groups differ in switching rate and failure-reactivity in both the QU and QC stages. Applying the classification rule to this problem reproduces the classification for 78.3% of subjects ($\kappa = 0.56$, $n = 120$). The share of optimal problems used by the original classification includes this problem.

- *Persistence.* Fifty of the 147 subjects played a later free-switching problem after the single-switching restrictions of Session 2 (29 noisy-type and 21 valuation-type). Valuation-type subjects switch more (0.25 vs. 0.09, $p < 0.001$) and react more to failures (0.30 vs. 0.11, $p < 0.001$). Among the 27 for whom Arm U never paid in this problem: 0.23 vs. 0.11 ($p = 0.019$) and 0.29 vs. 0.13 ($p = 0.003$).

QC-stage over-switching by type. The comparison in the main text uses the 141 classified subjects with at least one problem in which Arm U paid and, as elsewhere, defines QC-stage over-switching on the QC stage of the last problem in which Arm U paid. In a logit of QC-stage over-switching on a noisy-type indicator and the QU-stage switch rate, the coefficient on the indicator is -1.34 ($p = 0.012$) and that on the switch rate is 1.06 ($p = 0.47$). The share of optimal problems used by the classification includes the QU stage of the last problem in which Arm U paid. Recomputing it without that problem changes the label of 17 subjects (69 noisy type and 78 valuation type); valuation-type subjects then make 6.1 extra QC-stage switches against 1.3 ($p < 0.001$), and the logit coefficient on the noisy-type indicator is -1.57 ($p < 0.001$).

7.11 Horse Race between Plan Models and Models without a Plan

Data. Decisions are rounds $t = 1, \ldots, 29$; $y_{it} = 1$ if subject i switches arms between t and $t+1$. QU-stage decisions are those before Arm U first pays, QC-stage decisions those after. Samples: the 147 Treatment 1 over-switchers (QU) and those among them who over-switch in the QC stage of their last problem in which Arm U paid (QC); a subject needs at least three problems with decisions in a stage. Covariates: consecutive failures in the current wave (reset by a reward or a switch), draws in the current wave, and the posterior that Arm U is good, $P_{it} = 0.5 \cdot 0.8^{f_{it}} / (0.5 \cdot 0.8^{f_{it}} + 0.5)$, with f_{it} the failures on Arm U so far in the problem ($P_{it} = 1$ once Arm U has paid).

Models. *Plan models* follow a single-switching plan: stay on Arm U until the number of failures on Arm U reaches a cutoff K, then move to Arm F and stay there. In the QC stage the plan is to stay on Arm U and there is no cutoff. The five plan models differ in how subjects deviate from the plan:

1. One random slip: the same probability of deviating whether the plan says to stay or to switch.

2. Two random slips: a probability of switching when the plan says to stay (the slip) and a probability of staying when the plan says to switch, at most 0.5.

3. Failure-driven slip: two slips, with a slip that rises with the number of consecutive failures in the current wave, at the same rate on both arms.

4. Failure-driven slip per arm: as the previous model, but the slip rises at a different rate on each arm.

5. Wave-length slip per arm: two slips, with a slip that rises with the number of draws in the current wave, at a different rate on each arm.

Models without a plan give each subject a baseline switching propensity α_i (on the logit scale), which depends on:

6. Consecutive failures: the number of consecutive failures in the current wave, with the same slope on both arms;

7. Consecutive failures per arm: the number of consecutive failures in the current wave, with a different slope on each arm;

8. Wave length per arm: the number of draws in the current wave, with a different slope on each arm;

9. Posterior: the posterior that Arm U is good ($2P - 1$), with a different coefficient on each arm;

10. Posterior and consecutive failures: both the posterior and the number of consecutive failures in the current wave, with different coefficients on each arm.

A constant propensity, rejected in Section 4.2, is not included. In the QC stage, where the posterior equals one, models 9 and 10 are not estimated.

Specification. Let Λ be the logistic function and s_j the shift of problem j ($s_1 = 0$). Models without a plan set

$$\Pr(y_{it} = 1) = \Lambda\left(\alpha_i + x'_{it}\beta_i + s_j\right),$$

where x_{it} holds the covariates of models 6–10, interacted with the current arm when the slopes differ by arm. In plan models, the prescribed arm is Arm U while $f_{it} < K$ and Arm F once $f_{it} \geq K$ (in the QC stage, Arm U throughout), and

$$\Pr(y_{it} = 1 \mid K) = \begin{cases} \Lambda\left(\alpha_i + x'_{it}\beta_i + s_j\right) & \text{on the prescribed arm (the slip),} \\ \Lambda\left(\operatorname{logit}(1 - \varepsilon_i) + s_j\right) & \text{otherwise,} \end{cases}$$

where x_{it} is empty in models 1–2, consecutive failures in 3–4 and draws in the wave in 5, and $\varepsilon_i \leq 0.5$ is the probability of staying when the plan says to switch ($\varepsilon_i = \Lambda(\alpha_i)$ in model 1). In each problem $K \in \{1, \ldots, 30\}$ ($K = 30$: never leave Arm U) has probability $\pi_i(K) \propto \exp[-(K - \mu_i)^2/(2\sigma^2)]$ and is integrated out: a problem's likelihood is $\sum_K \pi_i(K) \prod_t \Pr(y_{it} \mid K)$. Bounds: baseline probabilities in $[0.01, 0.99]$, baseline slip ≤ 0.5, slopes in $[-3, 3]$ ($[-10, 10]$ for the posterior), $\mu_i \in [1, 30]$. Predicted probabilities are bounded away from 0 and 1 by 10^{-9}, only to avoid numerical errors.

Estimation and scoring. The parameters α_i, β_i, μ_i and ε_i are personal and estimated by maximum likelihood. From a subject's third problem in a stage on, each problem is predicted with parameters estimated on their earlier problems only; problems with fewer than three decisions are not scored. The shifts s_j and the dispersion σ are common to a group and held fixed: s_j from a model with a personal baseline and problem dummies only, applied to every model, and σ from a fit of each plan model to the whole group with a

common center. Accuracy is the geometric mean of the probability assigned to the observed choices, pooled over the scored decisions of a group. The best plan model of a group and its best model without a plan are the models of each family with the highest accuracy in that group. A subject is best described by a plan if a plan model has the highest out-of-sample log-likelihood among the ten. The *gap test* (the rank test of the main text) compares, between groups, each subject's out-of-sample log-likelihood per decision of their best plan model minus that of their best model without a plan (one-sided Mann–Whitney test). The *count test* compares, between groups, the share of subjects best described by a plan (one-sided Fisher exact test). Confidence intervals resample subjects (4,000 draws) with the best models held fixed.

Results. Table 32 reports the main results. The anchor columns follow the anchors' treatment-based labels, while each anchor's models use the group-level quantities of the group to which the classification assigns it; one noisy-execution anchor has too few problems to be scored.

Table 32: Plan versus no plan, by group

	QU stage		QC stage		Anchors (QU)	
	Noisy	Valuation	Noisy	Valuation	Noisy	Valuation
Subjects	80	63	18	15	9	8
Best described by a plan (%)	61	44	67	53	89	38
Accuracy, best plan model (%)	66.6	52.5	62.5	59.4	69.2	55.4
Accuracy, best model without a plan (%)	61.0	50.9	59.8	54.9	65.6	56.3
Difference significant at 5%	yes	no	no	no	no	no

Notes: Best plan model: two random slips (one random slip for noisy-execution anchors). Best model without a plan: posterior (QU), consecutive failures (QC and anchors). QU stage: gap test $p = 0.002$, count test $p = 0.033$. Anchors: count test $p = 0.043$.

Robustness. Table 33 repeats the QU comparison under two changes. (A) The group-level shifts and dispersion of the cutoff are estimated without the subjects being scored (10 folds of subjects within each group). (B) Each predicted problem is assigned to a group by a classification that leaves it out: the three markers are recomputed without that problem and the subject is reassigned to the nearest centroid. This changes the group of 25 of the 337 scored problems. Seventeen problems cannot be classified this way, because they are the subject's only problem in which Arm U never paid, from which two of the markers are computed; the table therefore also reports the main classification on the same problems.

Table 33: Plan versus no plan, QU stage: robustness

	Best described by a plan		Accuracy: plan vs. no plan (%)		Gap test
	Noisy	Valuation	Noisy	Valuation	p
Main specification	49/80	28/63	66.6 vs. 61.0*	52.5 vs. 50.9	0.002
(A) Group parameters without scored subjects	49/80	29/63	66.5 vs. 61.0*	52.3 vs. 50.8	0.002
(B) Same problems, main classification	43/77	28/60	67.7 vs. 64.1*	52.6 vs. 52.6	0.015
Leave-one-out classification, each problem	49/84	29/72	66.5 vs. 62.9*	53.1 vs. 53.7	0.002
Leave-one-out, subjects with one label	35/65	24/53	68.5 vs. 64.6*	52.1 vs. 52.5	0.024

Notes: *: 95% confidence interval of the difference excludes zero. In (B), a subject whose problems receive different labels contributes to each group with the corresponding problems ("each problem"), or is dropped ("subjects with one label"). Count test p-values, in row order: 0.033, 0.050, 0.19, 0.018 and 0.23.

Rare slips. Without restrictions, the slip of the two-slip plan model exceeds 0.25 for 76% of valuation-type and 11% of noisy-type subjects, and exceeds 0.10 for 98% and 74%. Table 34 and Figure 8 cap the slip probability of all plan models, with every other element of the main specification unchanged. For noisy-type subjects the advantage of the plan hardly moves: it stays between 4.4 and 5.7 points even with slips capped at 0.10, a bound that binds for three quarters of them. For valuation-type subjects it falls from 1.5 points to -2.4, and at the tightest cap the models without a plan predict them significantly better. The cap applies to a subject's own slip probability, before the problem-level shifts. In later problems these shifts lower switching for noisy-type subjects but hardly for valuation-type subjects, so with a cap of 0.25 the effective limit in those problems is about 0.09 for the former and 0.24 for the latter. This works against the result. When both probabilities are capped, the advantage of the plan for noisy-type subjects also shrinks (-3.0 points at 0.10 and $+3.5$ at 0.25; significant only from 0.35), while the difference between the groups grows (gap test $p \leq 0.001$ at every cap). Noisy-type subjects rarely leave the arm the plan prescribes, but they do not switch at a precise moment: their probability of staying when the plan says to switch has a median of 0.5, its upper bound.

Relative distance to the anchor centroids. Of 10,000 random splits of the same subjects into groups of 80 and 63, 0.18% give a gap test as small as the observed one. Holding the switching rate fixed, the relative distance predicts the gain from the plan model ($p = 0.013$) and the switching rate does not ($p = 0.27$), in a rank regression of each subject's gain from the plan model on the relative distance to the two centroids and on the switching rate over all QU-stage decisions. With the switching rate measured only on problems that are never scored, $p = 0.001$ for the relative distance and $p = 0.08$ for the switching rate. The relative

Table 34: Plan versus no plan with rare slips, QU stage

Maximum slip	Best described by a plan		Plan minus no-plan accuracy (points)		Gap test
probability	Noisy	Valuation	Noisy	Valuation	p
0.10	49/80	16/63	$+4.4^*$	-2.4^*	< 0.001
0.15	49/80	18/63	$+5.2^*$	-2.0	< 0.001
0.20	51/80	20/63	$+5.5^*$	-0.6	< 0.001
0.25	50/80	21/63	$+5.7^*$	$+0.3$	< 0.001
0.30	50/80	23/63	$+5.7^*$	$+0.8$	< 0.001
0.35	50/80	26/63	$+5.7^*$	$+1.1$	0.001
0.40	50/80	26/63	$+5.7^*$	$+1.4$	0.002
No cap	49/80	28/63	$+5.6^*$	$+1.5$	0.002

Notes: *: 95% confidence interval excludes zero. The cap applies to the subject's own slip probability, before the problem-level shifts.

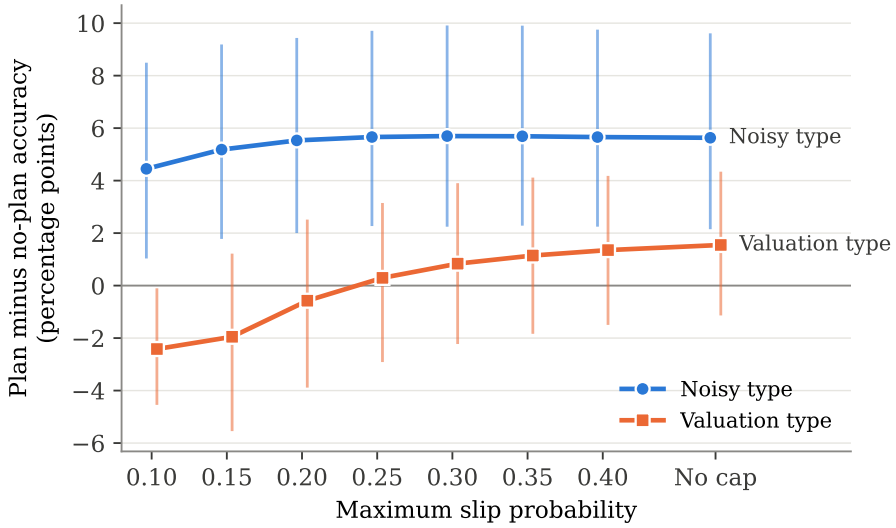


Figure 8: Plan versus no plan, by maximum slip probability

Notes: Difference in out-of-sample accuracy (percentage points) between the best plan model and the best model without a plan, QU stage, with 95% bootstrap intervals over subjects. The maximum slip probability caps the probability of switching when the plan says to stay; all other parameters as in the main specification. 80 noisy-type and 63 valuation-type subjects.

distance and the switching rate are highly correlated (0.86).

Parameters. Table 35 reports the parameters fitted on all problems, with the problem shifts and the dispersion of the cutoff estimated on both groups together so that the groups are comparable. Valuation-type subjects slip about twice as often (median 0.36 vs. 0.16 in the two-slip plan model; $p < 0.001$) and have lower cutoff centers (medians 1.0 vs. 4.0 failures, Mood's median test $p = 0.008$; 1.9 vs. 8.1 in the one-slip model, $p = 0.001$). When slips can grow with failures or with the length of the wave, the cutoff centers no longer differ significantly, and in the per-arm models the slips of valuation-type subjects grow much faster on Arm U ($p < 0.001$). In the models without a plan, the baseline propensity of valuation-type subjects is more than twice as high (medians 0.42 to 0.48 vs. 0.16 to 0.24; $p < 0.001$ in every model), while the slopes do not differ at the 5% level.

Table 35: Fitted parameters by group, QU stage

Model	Parameter	Noisy type	Valuation type	p
1. One random slip	Slip (switch when the plan says stay)	0.26	0.39	< 0.001
	Cutoff center (failures)	8.1	1.9	0.001
2. Two random slips	Slip (switch when the plan says stay)	0.16	0.36	< 0.001
	Stay when the plan says switch	0.50	0.40	< 0.001
	Cutoff center (failures)	4.0	1.0	0.008
3. Failure slip, one slope	Slip (switch when the plan says stay)	0.09	0.28	< 0.001
	Stay when the plan says switch	0.50	0.41	< 0.001
	Cutoff center (failures)	5.0	1.0	0.207
	Wave failures	0.13	0.22	0.686
4. Failure slip, per arm	Slip (switch when the plan says stay)	0.08	0.24	< 0.001
	Stay when the plan says switch	0.50	0.47	< 0.001
	Cutoff center (failures)	4.2	3.2	0.793
	Wave failures, Arm U	0.05	0.61	< 0.001
	Wave failures, Arm F	0.18	0.32	0.160
5. Duration slip, per arm	Slip (switch when the plan says stay)	0.08	0.17	< 0.001
	Stay when the plan says switch	0.50	0.48	0.001
	Cutoff center (failures)	2.8	1.0	0.207
	Wave length, Arm U	0.02	0.85	< 0.001
	Wave length, Arm F	0.16	0.22	0.686
6. Wave failures, one slope	Baseline propensity	0.19	0.43	< 0.001
	Wave failures	0.09	0.00	0.054
7. Wave failures, per arm	Baseline propensity	0.17	0.42	< 0.001
	Wave failures, Arm U	0.24	0.24	1.000
	Wave failures, Arm F	0.10	0.02	0.207
8. Wave length, per arm	Baseline propensity	0.16	0.42	< 0.001
	Wave length, Arm U	0.21	0.26	0.460
	Wave length, Arm F	0.08	0.02	0.111
9. Posterior, per arm	Baseline propensity	0.24	0.48	< 0.001
	Posterior, Arm U	-0.63	-0.33	0.284
	Posterior, Arm F	0.41	0.77	0.686
10. Posterior and wave failures, per arm	Baseline propensity	0.19	0.44	< 0.001
	Wave failures, Arm U	0.14	0.19	0.686
	Wave failures, Arm F	0.21	0.09	0.207
	Posterior, Arm U	-0.44	-0.02	0.686
	Posterior, Arm F	1.11	1.14	0.943

Notes: Medians across subjects of parameters fitted on all Treatment 1 problems of the QU stage; 84 noisy-type and 63 valuation-type subjects. The problem shifts and the dispersion of the cutoff are estimated on both groups together. Slips, the probability of staying when the plan says to switch (at most 0.5) and the baseline propensity are probabilities; slopes are on the logit scale. p: Mood's median test.

7.12 Experiment Instructions

Decision Environment

This experiment consists of several decision problems. Each problem is made up of **30 rounds**.

Your Task

In each round of a problem, you will choose between two options: **Jar A** or **Jar B**. A ball will be drawn from the jar you chose and your payoff will depend on the color of the ball (more details on the next page).

At the end of the experiment, the computer will randomly select one of the problems to determine your payment (all problems have an equal chance of being selected). You will be paid based on the result of that one selected problem only.

Go to Comprehension Questions

Which of the following statements is true?

- ○ I only choose a jar in one round per problem.
- ○ In each round, I will have the option to select Jar A or Jar B.

How many problems will determine your final payment?

- ○ One randomly selected problem (all problems have an equal chance of being selected).
- ○ You won't get paid based on your choice in problems.

Next

Decision Environment

You will play **5** problems, each consisting of 30 rounds.

In each round, you will choose either Jar A or Jar B. Each jar contains 10 balls.

A ball will then be randomly drawn from the jar you selected (each ball has an equal chance of being drawn):

- If the ball is **black**, you receive 50 points.
- If the ball is **white**, you receive 0 points.
- Your total score for the problem is the sum of points from all rounds.
- 100 points are equivalent to $1 payoff.

The **ball is returned to the jar** at the end of each round, so the jar's composition stays the same throughout the problem.

More information about the composition of each jar will be provided on the next pages.

Go to Comprehension Questions

How are the points allocated each round?

- I receive 0 points if the ball is black and 50 points if it is white.
- I receive 50 points if the ball is black and 0 points if it is white.
- I receive 50 points randomly regardless of the color.

How will your total score for a problem be determined?

- Your total score for the problem is the sum of points for all rounds.
- Your total score for the problem is the sum of points for some rounds.

After a ball is drawn, what happens to it?

- The ball is always returned to the jar at the end of the round.
- The ball is not returned to the jar at the end of the round.
- The ball is sometimes returned to the jar at the end of the round.

How do points translate to your payoff?

- 10 points corresponds to $1 of payoff
- 100 points correspond to $1 of payoff
- 1000 points correspond to $1 of payoff

Next

Time remaining: **2:22**

Composition of the jars:

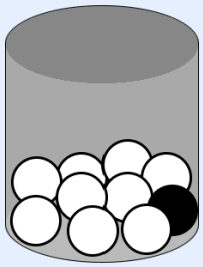
Jar B

Jar B is always good.

Good Jar B: Contains 9 white balls and 1 black balls. The probability of drawing a black ball (getting 50 points) is 10%.

- Notice that this is different from a good Jar A, which has 20% chance of returning a reward.

The figure below illustrates the Jar B:



Jar B (always good)

Go to Comprehension Questions

What is the chance that Jar B is good?

- 100%, since Jar B is always good.
- You do not know the chance that Jar B is good.

What is the chance of drawing a black ball from the Jar B?

- 10% (contains 1 black balls).
- 20% (contains 1 black balls).
- 0% (contains 1 black balls).

Next

.

Time remaining: **2:23**

Composition of the jars:

Jar A

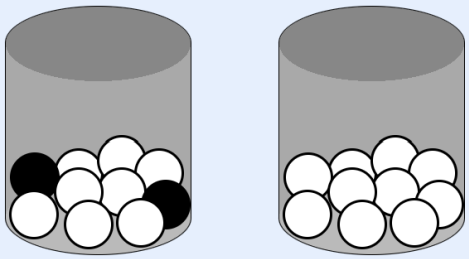
Jar A will be randomly chosen from a group of 2 jars, each with 10 balls. Among these, 1 is good and 1 is bad. Each jar has an equal chance of being selected.

- **At the start of each problem**, the computer randomly selects one of the 2 jars.
- Since 1 of the 2 possible jars is good, there is a 50% chance that Jar A is good and a 50% chance that it is bad.
- Jar A remains fixed for all 30 rounds of that problem.
- At the start of a new problem, a new Jar A is selected.
- The computer will not tell you whether the jar is good or bad.

Jar A Types

Good Jar A: Contains 8 white balls and 2 black balls. The probability of drawing a black ball (getting 50 points) is 20%.
Bad Jar A: Contains 10 white balls. The probability of drawing a black ball (getting 50 points) is 0%.

The figure below illustrates the difference between a good jar A and a bad jar A:



Good Jar A Bad Jar A

Go to Comprehension Questions

What is the chance that Jar A is good?

- Since 1 of the 2 jars is good, there is a 50 percent chance that Jar A is good and a 100 percent chance that Jar A is bad.
- You do not know the chance that Jar A is good.

What is the chance of drawing a black ball from the different types of Jar A?

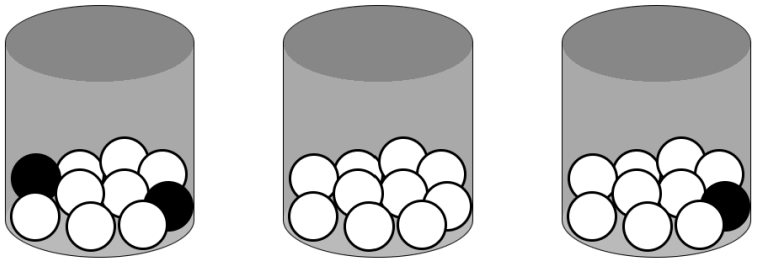
- 20% for good jar (contains 2 black balls) and 0% for bad jar (contains 0 black balls).
- 20% for good jar (contains 2 black balls) and 20% for bad jar (contains 0 black balls).
- 0% for good jar (contains 2 black balls) and 20% for bad jar (contains 0 black balls).

How often do you get a new Jar A?

- You get a new Jar A at the start of each round.
- You get a new Jar A at the start of each problem.
- Jar A is selected once and stays the same for all problems and rounds.

Time remaining: **2:23**

Summary of jars:



Good Jar A | Bad Jar A | Jar B (always good)

Keep in Mind!

- After observing a white ball from Jar A, you cannot know if the jar is bad (with no black balls) or good. (where there's still a chance to draw a black ball in subsequent rounds).
- Rewards can only come from good jars, as only good jars contain black balls.

Go to Comprehension Questions

If you see a white ball from Jar A, which statement is true?

- Jar A is certainly bad, so a black ball will never be drawn.
- Jar A is certainly good and may yield a black ball in future rounds.
- Drawing a white ball does not provide certainty about whether Jar A is good or bad.

If you see a black ball from Jar A, which statement is true?

- Jar A is certainly bad, so a black ball will never be drawn.
- Jar A is certainly good and may yield a black ball in future rounds.
- Drawing a white ball does not provide certainty about whether Jar A is good or bad.

Next

Choice Screen Guide

When you make choices, the screen will remind you of how each jar was selected and how likely it is to give a reward if good. Below is an example:



The computer will show how many rounds you have chosen Jar A without drawing a black ball:

Draws without reward:

The message below tells you how the jar was selected and how likely it is to give a reward if good:
Reminder: Jar A was randomly chosen from a group with **50%** good jars.

A *good Jar A* returns a reward with probability **20%**.

You don't know yet whether Jar A is good or bad.



The message below reminds you that Jar B is always good and how likely it is to give a reward.
Jar B is ***Good***.

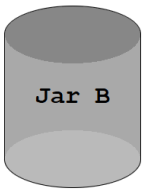
It returns a reward with probability **10%**.

If you find a black ball from Jar A, you then know that the jar is good. The information message will update.
The picture below shows an example *after Jar A returned a black ball.*



The message below updates if Jar A is revealed to be good.
Update: You now know that Jar A is ***good***.

It returns a reward with probability **20%**.



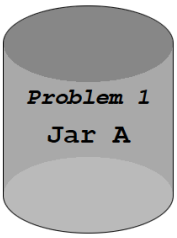
The message below reminds you that Jar B is always good and how likely it is to give a reward.
Jar B is ***Good***.

It returns a reward with probability **10%**.

Next

Round 1

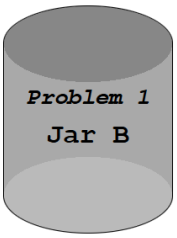
Please click on the jar you want to draw from this round:



Reminder: This Jar was randomly chosen from a group with **50%** good jars.

A *good arm A* returns a reward with probability **20%**.

You don't know yet whether Jar A is good or bad.



Jar B is ***Good***.

It returns a reward with probability **10%**.

New Problem

A new Jar A has been selected from a set of 1 good jar and 1 bad jar.
Jar B is good.

When you choose Jar A, you will see the ball drawn from this newly selected Jar A and be paid according to its color.
When you choose Jar B, you will see the ball drawn from this newly selected Jar B and be paid according to its color.

Click to Start the Problem.

Next

Treatment 2 Instructions

Shown after the subject completed all the Treatment 1 choices.

Part 2

Decision Environment

The environment in this part is identical to Part 1 with two key differences:

- Once you choose Jar B, you can not go back to Jar A for the rest of the problem. Only Jar B will be available until a new problem starts.

  When you select Jar B for the first time, a warning message will show up reminding you this action is not reversible!

  > If you choose Jar B now, you will not be able to choose Jar A until a new problem starts.
  > To choose Jar B, click on the button again.

- If you get a black ball (reward) from Jar A — so you know that Jar A is good — you will only be able to choose Jar A for the rest of that problem, since a good Jar A is more likely to draw a reward (20%) than Jar B (10%).

Reminders from Part 1

- Drawing a black ball gives you a 40 points reward. Drawing a white ball gives you 0 points.
- Jar A is good with probability 50% and gets renewed every problem.
- A good Jar A returns a reward with probability 20% (2 balls out of 10 balls).
- A bad Jar A returns a reward with probability 0 (0 black balls out of 10 balls).
- Jar B is always good and returns a reward with probability 10% (1 out of 10 balls).
- The ball drawn is always returned to the jar at the end of the round.

*In this part, you will play 5 decision problems of **30 rounds each**. If this part is selected for payoff, one of the problems will be randomly selected to define your compensation for this experiment.*

Go to Comprehension Questions

Which of the following statements is true?

- ○ Once I choose Jar B in a problem, only Jar B will be available for the remaining rounds of the same problem.
- ○ I can switch multiple times between Jar A and Jar B in the same problem.

Which of the following statements is true?

- ○ If a black ball is drawn from Jar A -revealing that Jar A is good-, only Jar A will be available for the remaining rounds of the problem.
- ○ If a black ball is drawn from Jar A -revealing that Jar A is good-, I can switch multiple times between Jar A and Jar B for the rest of the problem.

Next

Treatment 3 Instructions

Shown after the subject completed all the Treatment 2 choices.

Part 3

This part works similarly to the previous two parts, with one key difference: Before starting each problem, **you will be asked to report the number of rounds you would like to draw from Jar A** before switching to Jar B.

For all but the last problem, this report is not a final decision, meaning that you will proceed round by round, observe the outcome, and have the chance to decide when to switch to Jar B, regardless of your report. Your report is just a guide to help you think about your plan.

However, for the **last problem**, your report is definite:

- You will not play round by round.
- Instead, the computer will automatically choose Jar A for the number of rounds you specified.

- If a reward is found during those rounds, the computer will keep choosing Jar A until de the end of the problem and you will receive the corresponding rewards.
- If no reward is found within the specified rounds, the computer will switch to Jar B, and you will receive the corresponding Jar B rewards.

The **total number of problems** in this part is **unknown** to you. That means any problem could be the last one. Therefore, when reporting how many rounds you'd draw from Jar A, we recommend assuming each problem could be the last.

Go to Comprehension Questions

What will you be asked to report at the start of each problem?

- ○ The number of rewards you expect to receive.
- ○ Whether you prefer Jar A or Jar B overall.
- ○ The number of rounds you want to draw from Jar A before switching to Jar B.

When is your report about the number of rounds you want to draw from Jar A definite (automatically implemented by the computer)?

- ○ In every problem in this part.
- ○ Only for the last problem in this part.

How many problems are there in this part?

- ○ 2
- ○ 3
- ○ Unknown, any problem could be the last one.

Next

Treatment 4 (Block Choice) Instructions

Shown after the subject completed all the Treatment 1 choices.

Part 2

This part works similarly to the previous part, with one key difference: instead of choosing a jar every round, you will now choose a jar and **how many consecutive rounds to draw from it (the segment size)** before choosing again.

How Segments Work

At the beginning of each segment, you will make two decisions:

- **Which jar** to draw from (Jar A or Jar B).
- **How many consecutive rounds** to draw from that jar **(the segment size)**.

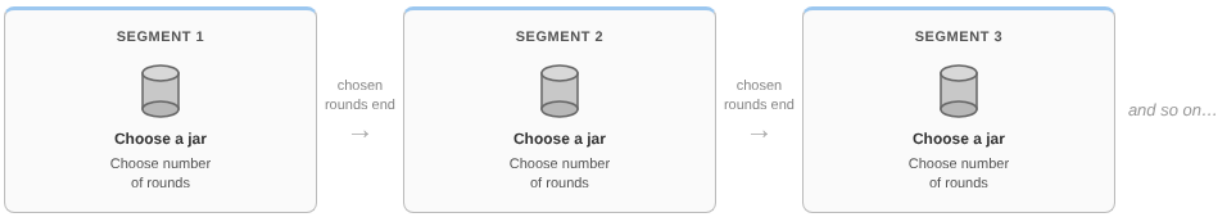
The computer will then draw from the selected jar for the number of rounds you specified. You will observe the outcome of each round one at a time, but you will not make another choice until the segment ends.

When the segment ends, you will again choose a jar and a segment size. **You may either choose the same jar as in the previous segment or switch to the other jar.** This continues until all 30 rounds of the problem are completed.

EXAMPLE

Suppose that in round 1 you choose Jar A for three rounds. The computer will draw from Jar A in rounds 1, 2, and 3. In round 4, you will choose a jar and a new segment size again.

How a problem is divided into segments



Special Case: What if Jar A draws a black ball?

A black ball immediately reveals that Jar A is good. If this happens:

- A message will confirm that Jar A is good.
- Your current segment will **end immediately**.
- In the next round, you will choose a jar and a new segment size again.

*In this part, you will play **2 decision problems** of **30 rounds each**. If this part is selected for payoff, one of the problems will be randomly selected to define your compensation for this experiment.*

Reminders

- Drawing a black ball gives you a **50 points** reward. A white ball gives **0 points**.
- Jar A is good with probability **50%** and is renewed every problem.
- A good Jar A returns a reward with probability **20%** (2 black balls out of 10).
- A bad Jar A returns a reward with probability **0%** (0 black balls out of 10).
- Jar B is always good and returns a reward with probability **10%** (1 black balls out of 10).
- The ball drawn is always returned to the jar at the end of the round.

Go to Comprehension Questions

How do you make choices in this part?

- ○ I choose a jar every single round.
- ○ I choose by segments. At the start of each segment, I choose a jar and how many rounds to draw from that jar.

If you choose Jar A for a segment of 4 rounds, what will happen after those 4 rounds?

- ○ I have to switch to Jar B for some rounds.
- ○ I can choose Jar A again or switch to Jar B for as many rounds as I want.

What happens when Jar A draws a black ball (reward) during your segment?

- ○ My segment continues as planned for the remaining segment rounds.
- ○ My current segment ends immediately. On the next round, I choose a jar and a new segment size, knowing that Jar A is good.

Next

Practice Problem

You will now play one **practice problem** (30 rounds). No payment is generated from this problem — it is for practice only.

The problems counting for payoff will start after the practice problem.

Click below to start the practice problem.

Next

Practice problem

Round 1

💡 **Practice guidance**

Just like in the previous parts, start by **clicking the jar** you want to draw from.

Click on the jar you want to draw from, then enter the number of rounds you want to keep this choice.



*Draws without reward: **0***

Reminder: This jar was randomly chosen from a group with **50%** good jars.

A *good Jar A* returns a reward with probability **20%**.

You don't know yet whether Jar A is good or bad.



Jar B is ***Good***.

It returns a reward with probability **10%**.

Your score so far: **0** points.

Practice problem

Round 1

Click on the jar you want to draw from, then enter the number of rounds you want to keep this choice.



*Draws without reward: **0***

Reminder: This jar was randomly chosen from a group with **50%** good jars.

A *good Jar A* returns a reward with probability **20%**.

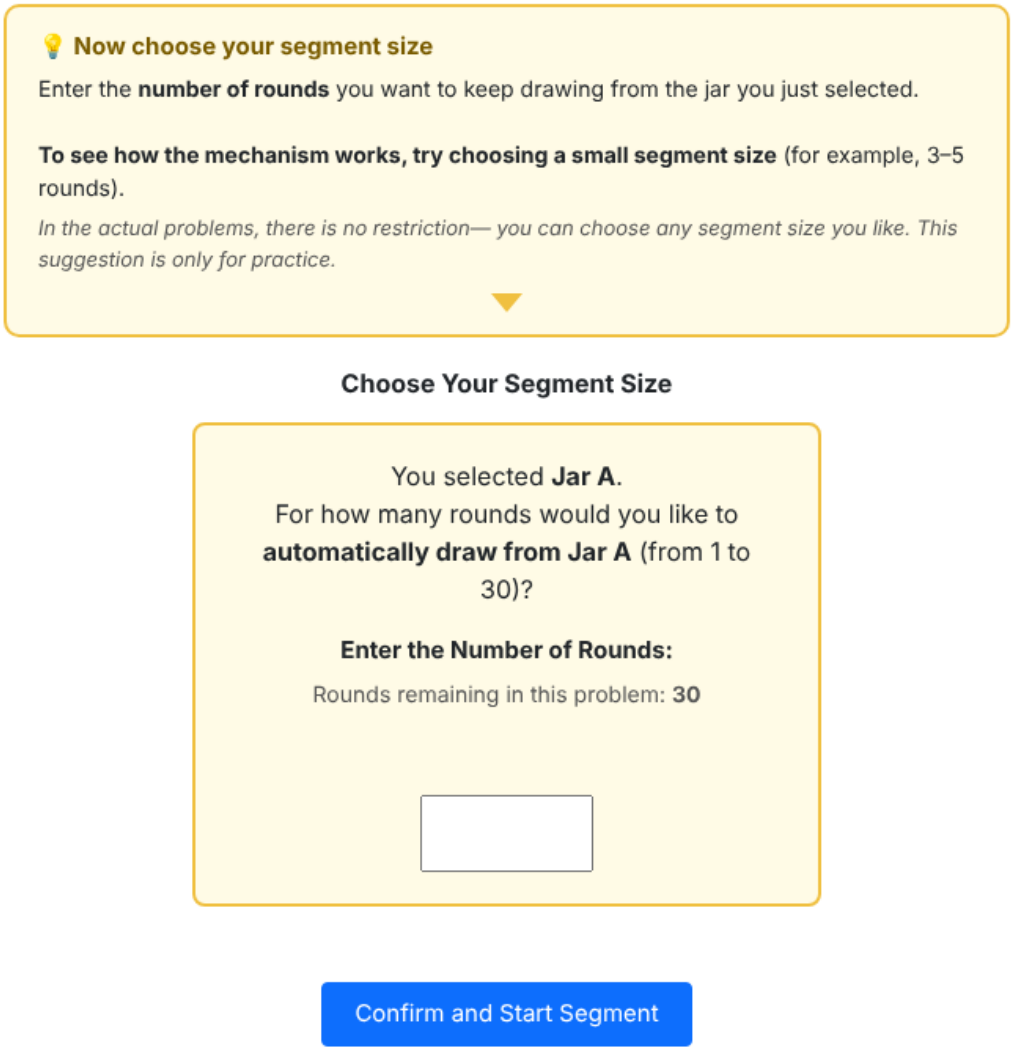
You don't know yet whether Jar A is good or bad.



Jar B is ***Good***.

It returns a reward with probability **10%**.

Your score so far: **0** points.


💡 Now choose your segment size
Enter the number of rounds you want to keep drawing from the jar you just selected.
To see how the mechanism works, try choosing a small segment size (for example, 3–5 rounds).
In the actual problems, there is no restriction— you can choose any segment size you like. This suggestion is only for practice.
Choose Your Segment Size
You selected Jar A.
For how many rounds would you like to automatically draw from Jar A (from 1 to 30)?
Enter the Number of Rounds:
Rounds remaining in this problem: 30
Confirm and Start Segment

Treatment 5 (Ex Ante Rules) Instructions

Shown after the subject completed all the Treatment 4 choices.

Part 3

Decision Environment

This is the last part of the experiment.

You will not make choices round by round, but the computer will choose on your behalf according to a rule you will select. You will only observe the final score.

You specify a switching rule for two phases: **Phase 1** when Jar A has not given a reward, and **Phase 2** if Jar A returns a reward.

This part works as follows: You specify the rules, the computer runs them in the background and shows you the final score. You will not observe anything else.

Reminder of the Environment

- Drawing a black ball gives you a **50 points** reward. Drawing a white ball gives you 0 points.
- Jar A is good with probability **50%** and gets renewed every problem.
- A good Jar A returns a reward with probability 20% (2 balls out of 10 balls).
- A bad Jar A returns a reward with probability 0 (0 black balls out of 10 balls).
- Jar B is always good and returns a reward with probability 10% (1 out of 10 balls).
- The ball drawn is always returned to the jar at the end of the round.

Next

Phase 1 — Switching Rules

Before you choose your Phase 1 rule, please review the four rules below once more and answer the questions at the bottom of the page to confirm you understand them.

Rule 1 — Switch after Consecutive White-Ball Draws

The computer switches jars after a chosen number of consecutive white-ball draws. You choose how many consecutive white balls to observe before switching from Jar A and from Jar B — the two thresholds can differ.

The computer starts on Jar A and alternates between the two jars whenever the threshold of white-ball draws for the current jar is reached. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: Suppose you choose a threshold of 3 for Jar A and 5 for Jar B. The computer starts on Jar A. After 3 consecutive white balls it switches to Jar B. After 5 consecutive white balls from Jar B it switches back to Jar A, and so on. If Jar A returns a black ball, Phase 2 begins.*

Rule 2 — Switch after Consecutive White-Ball Draws (Adaptive Thresholds)

Like Rule 1 (Consecutive White-Ball Draws), the computer switches jars after consecutive white-ball draws. The difference is that *each time the computer returns to a jar* you choose a new white-ball threshold for that jar.

The computer starts on Jar A and switches jars whenever the current threshold is reached. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: At the beginning you may choose to switch after 6 white-ball draws from Jar A, then switch back after 4 from Jar B. If the computer returns to Jar A, you may choose 2 for that visit, and 3 for the next visit to Jar B. You can choose different thresholds for every visit.*

Rule 3 — Single Switching

You choose how many rounds the computer will explore Jar A before permanently switching to Jar B. If Jar A returns a reward before the switch, the computer stays on Jar A for the rest of the problem (In Phase 2, the computer always chooses Jar A).

> *Example: If you choose to explore Jar A for 5 rounds, the computer draws from Jar A for rounds 1–5. If a black ball appears in any of those rounds, it keeps drawing from Jar A for the rest of the problem. If no black ball appears in rounds 1–5, the computer moves to Jar B for rounds 6–30.*

Rule 4 — Switch in Cycles

You choose how many rounds the computer stays on each jar before automatically switching to the other jar, regardless of the outcomes.

The computer starts on Jar A and switches jars after the chosen number of rounds. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: If you choose to switch every 5 rounds, the computer draws from Jar A for rounds 1–5, then from Jar B for rounds 6–10, then from Jar A for rounds 11–15, and so on.*

Comprehension Questions

Under which rule can the computer switch from Jar A to Jar B **only once**, and cannot switch back to Jar A?

- ○ Single switching
- ○ Switch after Consecutive White-ball Draws
- ○ Switch in Cycles

Under which rule does the computer switch between jars after a chosen number of consecutive white-ball draws?

- ○ Single switching
- ○ Switch after Consecutive White-ball Draws
- ○ Switch in Cycles

Under which rule do you switch between jars every chosen number of rounds, regardless of the outcomes?

- ○ Single switching
- ○ Switch after Consecutive White-ball Draws
- ○ Switch in Cycles

Next →

Phase 1 — Choose Your Rule

Please select one of the following rules for **Phase 1** (the rounds in which Jar A has not yet returned a reward, so you do not know its quality). You will choose the specific numbers on the next page.

Rule 1 — Switch after Consecutive White-Ball Draws

The computer switches jars after a chosen number of consecutive white-ball draws. You choose how many consecutive white balls to observe before switching from Jar A and from Jar B — the two thresholds can differ.

The computer starts on Jar A and alternates between the two jars whenever the threshold of white-ball draws for the current jar is reached. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: Suppose you choose a threshold of 3 for Jar A and 5 for Jar B. The computer starts on Jar A. After 3 consecutive white balls it switches to Jar B. After 5 consecutive white balls from Jar B it switches back to Jar A, and so on. If Jar A returns a black ball, Phase 2 begins.*

◯ **Select Rule 1**

Rule 2 — Switch after Consecutive White-Ball Draws (Adaptive Thresholds)

Like Rule 1 (Consecutive White-Ball Draws), the computer switches jars after consecutive white-ball draws. The difference is that *each time the computer returns to a jar* you choose a new white-ball threshold for that jar.

The computer starts on Jar A and switches jars whenever the current threshold is reached. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: At the beginning you may choose to switch after 6 white-ball draws from Jar A, then switch back after 4 from Jar B. If the computer returns to Jar A, you may choose 2 for that visit, and 3 for the next visit to Jar B. You can choose different thresholds for every visit.*

◯ **Select Rule 2**

Rule 3 — Single Switching

You choose how many rounds the computer will explore Jar A before permanently switching to Jar B. If Jar A returns a reward before the switch, the computer stays on Jar A for the rest of the problem (In Phase 2, the computer always chooses Jar A).

> *Example: If you choose to explore Jar A for 5 rounds, the computer draws from Jar A for rounds 1–5. If a black ball appears in any of those rounds, it keeps drawing from Jar A for the rest of the problem. If no black ball appears in rounds 1–5, the computer moves to Jar B for rounds 6–30.*

◯ **Select Rule 3**

Rule 4 — Switch in Cycles

You choose how many rounds the computer stays on each jar before automatically switching to the other jar, regardless of the outcomes.

The computer starts on Jar A and switches jars after the chosen number of rounds. If Jar A returns a black ball, the computer immediately moves to Phase 2.

> *Example: If you choose to switch every 5 rounds, the computer draws from Jar A for rounds 1–5, then from Jar B for rounds 6–10, then from Jar A for rounds 11–15, and so on.*

◯ **Select Rule 4**

Next →