Opportunistic exploration: Humans consider not just whether, but also when, to sample unfamiliar options

Jack Dolgin[1], Bettina Bustos[2], Robert C. Wilson[3], & Wouter Kool[1]

[1] Washington University in St. Louis
[2] University of Iowa
[3] University of Arizona

Author Note

Correspondence concerning this article should be addressed to Jack Dolgin, Somers Family Hall 1125, 1 Brookings Drive, St. Louis, MO 63105. E-mail: jdolgin@wustl.edu

Abstract

Big or small, many decisions incorporate the tradeoff between exploration and exploitation—whether to take advantage of what we know to be good, or to take a chance on something new. Recent research suggests we make this choice via a combination of stochasticity and directedness, and that the directedness involves prioritizing lesser seen options. Through a series of multi-armed bandit experiments, we extend this conceptualization of directed exploration to incorporate opportunistic choice in dynamic decision contexts. Participants chose between two bandits and, unlike in prior work, did not explore undersampled bandits more when more future trials remained, even though there was increased strategic value in learning about those choice options. Crucially, however, on each trial they chose one bandit by multiplying two randomly selected numbers, and the other bandit by adding the two numbers. We found that people seized the context to opportunistically explore, such that they were more likely to pass on the multiplication bandit for a hard problem when more subsequent trials remained. The results echo recent machine learning work on "opportunism" but in humans and suggest directed exploration reflects not just whether, but also when to explore.

Keywords: Exploration, Exploitation, Opportunism, Planning, Decision Making, Reinforcement Learning

Opportunistic exploration: Humans consider not just whether, but also when, to sample unfamiliar options

Introduction

Every day we perpetually size up and decide between options, and for many of those choices we have more familiarity with some options than others. In the case of a sports coach, she may have to decide which of her players, varying in experience, to put on the field. She could stick with those who are battle-tested and whose output is predictable; alternatively, she could take a chance on a young prospect who, if given the chance, might in fact prove themselves to already be the best. This kind of choice, pitting the unfamiliar vs. the predictable, is a variant of a long-researched problem known as the explore-exploit dilemma. Findings have consistently pointed to humans' aptitude for adaptively negotiating these scenarios, thanks both to domain-general and context-specific strategies. Across contexts, people decide how much to explore using two strategies—directed and random exploration—that mirror strategies computer scientists use to model optimal behavior in similar paradigms (e.g., Gershman, 2018; Wilson, Bonawitz, Costa, & Ebitz, 2021). In specific circumstances, we also may tack on additional bespoke strategies. For example, when exploring certain options makes more options available downstream, we prioritize these activating or "empowering" options (Brändle, Stocks, Tenenbaum, Gershman, & Schulz, 2023).

Nevertheless, the strategies that exploration research has identified are only a fraction of the strategies people have been shown to use across all kinds of decision-making tasks. One of the key omissions from the exploration research has been the strategy of when, rather than just whether, to explore. Previous studies frame the explore-exploit choice as static, as if options are selected one at a time. However, the sports coach from before, for example, may not be contemplating only whether to explore (inserting the young prospect into the game). Her choice may also be about inserting the player when the game is likeliest to be out of hand (and the risks to exploring are low). Even if exploring is tempting at the beginning of the game, the coach can get much of the same benefit (learning how talented the player is) by waiting until the end to insert the prospect. In that spirit, this paper investigates whether, and if so how, people sequence when to explore. Specifically, we ask whether people leverage the learnt, time-varying cost structure of the environment to opportunistically put off or pounce on chances to sample unfamiliar options.

Of course, it is long known that people can and frequently do plan the order of their actions and decisions. Lashley (1951) catapulted the idea that there is temporal organization behind actions, and, taking the mantle, Miller, Galanter, and Pribram (1960) characterized much behavior as planned. They contended that in planned behavior, people serially execute pre-decided steps to service a hierarchy of sub-goals and goals. Researchers like Hayes-Roth and Hayes-Roth (1978) suggested one principle that guides how people order these steps. They pointed out that, in the everyday world, the quality of choices often fluctuates across moments, and that people often match when they select an option based on its idiosyncratic dips and crests, if they were predictable (alternatively, if not predictable, people would adjust on the fly, not necessarily committing to predefined steps). They referred to this idea of "timing" as "opportunistic planning."

Yet to date, only inconclusive evidence has linked planning—opportunistic or

otherwise—to exploration. One indirect source of evidence stems from work by Hills and Hertwig (2010). Their participants sampled from two payoff distributions, and the results showed they consistently did so in one of two ways. Some participants stuck to sampling one distribution for half the time, then switched over to the other distribution for the second half. Another set of participants continually alternated between both distributions. These explorations of the distributions could signal planning, because there was a relationship between consecutive samples, potentially meaning that participants thought out in advance how they would order their behavior. On the other hand, participants may very well have adopted these patterns as mere heuristics, myopically repeating or zig-zagging without forethought. This kind of interpretation question also arises in the work of optimal stopping problems. In these problems, similar to the casino card game Black Jack, participants keep sampling (exploration) until they decide to walk away with their last reward. Participants could perform complex calculations, considering how many trials remain, evaluating the expected reward on an average trial, factoring in the likely variability, and settling on a threshold for when to stop sampling. Alternatively, they might not plan out what threshold they will stop at, instead making their call somewhat spontaneously. Thus, it remains unresolved whether people actively plan out how they will explore.

Work at the intersection of exploration and the form of planning central to this paper—opportunism—is less settled still. One study, deploying six bandits, informed participants that a bandit would disappear if they did not regularly sample it; the authors found that this threat caused participants to select unappealing bandits more often (Navarro, Tran, and Baz, 2018). This increase in sampling could reflect opportunistic exploration, since a bandit is explored specifically because it is a particularly opportune time to do so (it is available now, but will not otherwise be in the future); however, this behavior might have little to do with optimizing when to explore and may more reflect loss aversion (Kahneman & Tversky, 1979). Another study (Schulz, Klenske, Bramley, and Speekenbrink, 2017) controlled for any preservation tendencies, but the authors did not find compelling evidence of opportunistic exploration. In this virtual boat-steering task, anticipating high downstream costs to explore did not encourage more exploration at the outset. It therefore remains unclear whether opportunism might characterize exploration.

To test for opportunistic exploration, we created a bandit task where the value of a moment for exploration, defined in terms of mental effort, was sometimes particularly cheap and at other times particularly costly. Specifically, on each trial participants chose one bandit by multiplying two randomly selected numbers, or the other bandit by adding the numbers. Participants sometimes faced difficult multiplication problems (e.g., 19 and 13), or in more opportune moments sampling the multiplication bandit required completing only a simple calculation (1 and 13). Our hypothesis was that while a more difficult math problem might make one bandit particularly more aversive than the other, this aversiveness should be amplified when there are more future opportunities to explore that same side, especially under the expectation that the cost of exploring that same side will likely be lower (i.e., easier math problems) on future trials. Likewise, an easy math problem will loom especially large, we expected, when one might be in the thick of a longer series of trials, because if one hopes to sample that side anyways, they may pounce on the opportunity before facing a more difficult math problem. In short, when there are more

future opportunities to explore, participants should become more sensitive to the relative cost between bandits.

In total, we ran a baseline arithmetic task and three novel two-choice bandit tasks. Besides being realistically complex, without obvious demand effects, our bandit task also examined the presence of two of the most common types of uncertainty-driven strategies. In directed exploration, options that are less well known are most likely to be chosen, all else equal. In random exploration, choices are made randomly, which indirectly results in less explored options being sampled. In our task, bandits differed not only in associated math difficulty, but also expected rewards and number of previous samples (informativeness). Thus, we measured opportunism, random exploration, and directed exploration as three mutually compatible strategies.

Baseline Experiment

The baseline task indexed the difficulty of addition and multiplication problems. These math problems subsequently appeared during bandit task trials, thereby indexing their opportunism.

Materials and Methods

Participants. We recruited all participants from the Washington University pool of undergraduate psychology students. The tasks were approved by the Washington University Institutional Review Board, built using jsPsych (de Leeuw, 2015), and completed online, remotely, and for class credit. We recruited 129 participants for the arithmetic task and analyzed data from 115 (57 female, 56 male, 2 other) after removing participants for correctly answering fewer than 80% of trials.

Experimental Design. The entire task was answering 74-75 addition problems and 74-75 multiplication problems. Each problem featured two numbers, between 1 and 24. To generate the problems, we constructed 16 sets of addition and multiplication number pairs, with each pair appearing in one of the 16 sets as addends and again in another of the 16 sets as multiplicands. Each set consisted of 74-75 multiplicand and 74-75 addend pairs that were otherwise determined pseudorandomly. The pairs together represented all combinations of numbers between 1 and 24 except the couple 2 and 2. We randomly assigned a set to each participant so that the participant's task was answering the 148-150 problems in that set. Within each participant we randomized the order the pairs of numbers appeared. On a given trial, participants had unlimited time to answer, and they responded by typing their answer on their keyboard; problems lasting longer than 75 seconds were not further analyzed. After each response, they rated the difficulty of the problem on a scale from 0-100. We asked participants ahead of time to do all the math in their head, meaning not using the Internet or pen and paper, for example, as assistance. We also asked participants after the experiment if they followed the instructions about doing the math in their head, and excluded participants who did not.

Results

Measures of difficulty all significantly correlated with one another. At the group level, self-rated difficulty correlated with response time $r(596) = 0.96$, $p < .001$ and with

accuracy $r(596)$ = -0.89, p < .001. Because of the particularly tight link between self-rated difficulty and response time, we focus on these two measures in the rest of the paper. As seen in Figure 1, each participant found multiplication problems more difficult than addition problems. However, Figure 2 reveals that this difference varied widely depending on the pair of numbers. For the median pair, multiplying the two numbers took 4.53 seconds longer than adding those same two numbers; however, that difference increased to 27.0 seconds in the case of one pair, 18 and 18, whereas on the other end of the spectrum, participants on average multiplied the pair 10 and 12 0.84 seconds *faster* than adding the two numbers. Therefore, the baseline task successfully produced a continuous and dispersed index of difficulty serviceable for the bandit experiments.

Bandit Experiments

We ran three bandit experiments in total. The first and third experiments were identical, and the second version was nearly the same. Key predictions for the second and third experiments were pre-registered, as shown in Tables 1-4, logged at osf.io/kb4xa/registrations, and further discussed in the Results section.

Materials and Methods

Participants.

Bandit Experiment 1. We recruited 129 undergraduate students for the first bandit task and analyzed data from 81 (46 female, 33 male, 2 other). We removed 8 participants with accuracies below 80% on math problems, 30 more who, despite instructions, reported that they did not complete all math problems entirely in their head, and another 10 who chose the same bandit more than 85% of the time.

Bandit Experiment 2. We recruited 241 individuals for the first preregistered bandit task and analyzed data from 150 (106 female, 41 male, 3 other). In line with our preregistration, we pruned 16 with accuracies below 80% on math problems, 66 more who, despite instructions, reported not completing all math problems entirely in their head, and another 9 who chose the same bandit more than 85% of the time.

Bandit Experiment 3. We collected data on a new set of 367 participants for the replication bandit task and analyzed the data of 200 (106 female, 88 male, 6 other). In line with our preregistration, we first excluded any participants who did not meet the exclusion criteria in Experiments 1 and 2. We pruned 34 with accuracies below 80% on math problems, 84 more who, despite instructions, did not complete all math problems entirely in their head, and another 23 who chose the same bandit more than 85% of the time. In addition, also in line with our preregistration, we excluded 0 participants who we suspected may have used a calculator, in spite of what they listed in the post-task questionnaire. These participants' data followed a similar pattern in choice time and math problem difficulty compared to those who listed post-task that they did use calculator help. Specifically, they chose a multiplication bandit at least 15 times when multiplying the two numbers was, on average in the pilot task, more than 12 seconds slower than adding them. Also, among those 15 or more choices, their average choice time was quicker than 10 seconds, suggesting they did not invest requisite effort into these difficult multiplication problems.

Experimental Design. We modeled the bandit experiments off Wilson et al.'s (2014) Horizon Task. The experiments were broken into 80 games, and each game was broken into a number of trials. At the start of a game, two vertical rectangles appeared, each sliced into the same number of pieces as the number of trials in that game (see Figure 3). We told participants that each vertical rectangle represented one bandit, and bandits were set up so they dispensed some average amount of rewards over the game that was variable from trial to trial. The rewards were always valued between 1 and 100 points.

We explicitly told participants that the bandits during one game would have zero relation with the rewards underlying subsequent bandits on either side. We also did not provide them any clues, other than the opportunity to sample from trial to trial, about which side was more likely to yield a greater reward. Under the hood, we randomly assigned one of the two bandits a generative average of either 60 or 40, and we randomly assigned the other bandit's generative mean as 20, 8, or 4 points different (these differences were 4-20 points lower if the first bandit was 60, and 4-20 points greater if the first bandit was 40). We also set the standard deviation of a bandit from its mean at 8 points.

The first four trials in each game were always forced choices. Specifically, participants were told whether to press the left arrow key, corresponding to the left rectangle, or the right arrow key, corresponding to the right one (the task ignored presses of the wrong key). In 50% of games, two choices were forced to each of the two rectangles, in a random order; in the other 50% of games one side was forced once and the other side three times (50% of these games involved three forced choices to the right side, 50% involved three forced choices to the left).

After the forced choices, participants were free to choose between bandits ("free choice" trials). Participants faced either 1 free choice trial (known as "short horizon" games) or 4-5 free choice trials ("long horizon" games). The singular difference between the first and third experiments vs. the second is that each long horizon was 4 trials for the first and third experiments but 5 trials for the second experiment. Each participant completed 40 short horizon games and 40 long horizon games, shuffled randomly. Note that at the onset of a game, participants could immediately see how many trials the game would last.

The novel and critical wrinkle in our design was that for the free choices (trials 5 and beyond), selecting a bandit entailed solving an arithmetic problem in lieu of pressing the corresponding arrow key. For each participant, the bandit on the left side of the screen in every game was associated with either addition or multiplication, and the bandit on the right side of the screen with the other of the two types of arithmetic (side and arithmetic type were randomized between participants). Specifically, we randomly presented one of 299 pairs of numbers from the arithmetic-only experiment on each trial 5 or beyond. The two numbers were always between 1 and 24, and participants decided whether to choose the bandit that required them to add these two numbers or to multiply them. For example, imagine a trial on which the numbers are 4 and 13. If the participant chose the multiplication bandit, then the participant needed to answer "17" (= 4 + 13), but if they chose the addition bandit, then they needed to answer "52" (= 4 x 13). If participants provided an answer that did not solve either problem, the trial ended and no points were provided. Moreover, we encouraged them to amass the most total points over the experiment, and further, we informed them that ranking among the top seven participants in points earned would amount to a $15 Amazon gift card on top of participation class

credit.

Analysis

We conducted our analyses in R (R Core Team, 2021), using the packages tidyverse (Wickham et al., 2019), arrow (Richardson et al., 2021), cowplot (Wilke, 2020), patchwork (Lin Pedersen, 2020), and gghalves (Tiedemann, 2020). The code for the online task and the analyses are available at https://github.com/jackdolgin/opportunistic_exploration. The results can also be explored interactively at https://jackdolgin.shinyapps.io/opportunistic_exploration. The only trial-level pruning occurred if a trial or any trial preceding it in the same game involved an incorrect math response (5.3% of trials were incorrect) or if the response time while answering was greater than three minutes.

Our statistical tests centered around a model used by studies with a similar task design (Zajkowski, Kossut, and Wilson, 2017; Feng, Wang, Zarnescu, and Wilson, 2021), with the addition of a difficulty parameter. The model, which takes the shape of a logistic choice curve, incorporates inputs that are different components of the choice. Our model is as follows:

$$p_{mult} = \frac{1}{1 + exp(\frac{R_{mult} - R_{add} + \alpha(I_{mult} - I_{add}) + D(M_{add} - M_{mult}) + B}{\sqrt{2}\sigma})}$$

In this equation, p_{mult} represents the probability of selecting the multiplication side, R_{mult} - R_{add} the difference in the mean points between the multiplication and addition bandits, and I_{mult} - I_{add} the difference in 'informativeness' between bandits. This latter difference equals 1 when the multiplication bandit is more informative (less sampled) than the addition bandit, -1 when the addition bandit is more informative, and 0 when both sides are evenly sampled. Thus, the greater the magnitude of the corresponding coefficient, α, the more sensitive participants are to bandit exposure. We interpreted α as a measure of directed exploration.

Because challenging math problems demand more cognitive effort (Hess & Polt, 1964; Dunn, Inzlicht, & Risko, 2019), which people experience as aversive (Kool, McGuire, Rosen, & Botvinick, 2010; Westbrook, Kester, & Braver, 2013), we expected participants to seek easier math problems and avoid more difficult ones. The difference M_{add} - M_{mult} represents how much more difficult it is to multiply the trial's numbers than to add them. We defined difficulty as the average time to solve the problem during the baseline experiment. Since pilot data revealed the distribution of RT differences to be skewed (see the bottom half of Figure 2), we entered the differences in RT between multiplication and addition as rank-ordered values, compared against the difference in RT of all 298 other pairs in the baseline experiment. The corresponding parameter, D, increased in magnitude when sensitivity to difficulty increased. We also recalculated D three other ways, resulting in four overall, slightly different models. First, we re-fit D in terms of raw RT differences, rather than performing rank-ordering. Second and third, we defined difficulty in terms of self-rated difficulty during the baseline; in the second model, we rank-ordered self-rated difficulty; in the third model, we used raw self-rated difficulty. All four approaches to quantifying D yielded similar fits, as RT and self-rated difficulty were correlated 0.96 in the

baseline task.

We standardized the difference in difficulty (regardless of how it was defined) and difference in reward so they spanned between -1 and 1 and therefore sat on the same scale as the information difference. The last two free parameters, B and σ, represent a bias for bandits on one side of the screen vs. the other and represent decision noise, respectively. This decision noise corresponds to how randomly a participant explored with respect to the other free parameters. We interpreted decision noise as a proxy for random exploration.

We fit the model separately for each free choice trial number for each horizon length for each participant using a maximum a posteriori estimation. As a result, there were six fits for each participant in Experiments 1 and 3 and seven for each participant in Experiment 2 (technically, there were four times as many fits, since we ran four models, one for each of the four difficulty parameters). We added an exponential prior (with length scale 20) for the temperature coefficient and a N(0, 20) Gaussian prior distribution for the information, difficulty, and side bias coefficients (Feng et al., 2021). In our first two experiments, we used a pair-wise t-test to assess whether coefficients significantly differed between the first free choices of short horizons vs. long horizons. Since the fits were not normally distributed, in our third experiment, we used a Wilcoxon signed-rank test in accordance with our pre-registration. A global optimizer from the RcppDE R package (Eddelbuettel, 2018) set bounds of -100 to 100 for the side bias, information, and difficulty coefficients, set a bound of 0 to 100 for the temperature coefficient, and ran up to 1,000 iterations per fit until the objective function had been minimized.

Results

Using generalized linear models, we performed several manipulation checks on our data to ensure their sufficient quality. As expected, greater expected reward, as measured by the mean standardized points on previous pulls to the bandit, significantly increased the odds a bandit would be picked ($\beta = 1.17$, 95% CI [1.16, 1.19], z = 20.70, p < .001). Conversely, the more difficult the arithmetic solution required to pull a bandit, as measured by the response time of that math problem in the baseline experiment, the lower the odds that side would be picked ($\beta = 0.49$, 95% CI [0.48, 0.50], z = -73.50, p < .001). Still, though multiplication problems in general were more difficult than their addition counterparts ($\beta = 17.52$, 95% CI [15.10, 19.43], $t(596) = 10.03$, $p < .001$), participants picked the multiplication side 39.8% of the time, indicating that people engaged with the experiment and did not myopically choose only the easiest math problem.

Our key question was whether people leveraged the learnt cost structure of the environment to opportunistically put off or pounce on chances to explore. Such a prediction would result in more sensitivity to costs at the start of long horizons compared to short horizons, when there are no future opportunities to explore and therefore is no putting off or pouncing to be done. The results in Figure 4, particularly for Experiments 1 and 2, align with this hypothesis. When participants had seen both bandits an equal number of times via forced choice, when the rewards those bandits had yielded so far had been comparable (on average), and when the math problems for one bandit were much harder than those for the other bandit, there still seemed to be a discernible avoidance of the multiplication side at the start of a long horizon, compared to during a short horizon. We formally tested

sensitivity to difficulty via the D coefficient in the model described in the Analyses section.

In Experiment 1, the D coefficient (Figure 5) was significantly larger at the start of long horizons than at the start of short horizons (M_d = 5.64, 95% CI [3.92, Inf], $t(80)$ = 5.45, $p < .001$). In Experiment 2, our pre-registered test for the difference in the difficulty parameter, calculated in terms of rank-ordered RT, was marginally significant (M_d = 1.42, 95% CI [-.04, Inf], $t(149)$ = 1.61, p = .055). However, as mentioned in the Methods section, there were three other equally plausible ways of calculating difficulty, and those approaches each reached significance (see Table 1; p-values all less than .01). Moreover, all four approaches reached significance when we implemented a Wilcoxon signed-ranked test, reflecting the non-normal shape of the parameter distribution. Therefore, in Experiment 3 we increased our sample size, pre-registered all four methods for calculating difficulty, and measured significance with a Wilcoxon signed-ranked test. As seen in Table 1, the difference in the difficulty parameter was significant across all measurement approaches (p-values of .001, .049, .001, and .044). These results suggest greater sensitivity to bandit math difficulty at the start of long horizons compared to short horizons.

Alongside opportunism, we also examined the mutually compatible presences of directed and random exploration strategies. These strategies predict increased long horizon exploration and an increased short horizon exploitation of reward, similar to the results in Experiment 2 depicted in Figure 4. Specifically, in Experiment 2 greater reward for one bandit on free choice 1, short horizon made that bandit particularly more compelling than when the bandit was equally rewarding on free choice 1, long horizon. To formally test the presence of these exploration strategies, we measured whether their corresponding model coefficients—for informativeness and σ for decision noise—were greater for the first choice on long horizon trials compared to short horizon trials.

Across the three studies, we found inconclusive evidence for such information and noise increases (see Tables 2 and 3 and Figures 6 and 7). In Experiment 1, how we defined difficulty influenced whether the informativeness coefficient was significantly greater at the start of the long horizon (p-values ranged between .032 and .055) (see Table 2). Meanwhile, we observed no difference in the decision noise coefficient between short and long horizons (M_d = 0.10, 95% CI [-Inf, 1.08], $t(80)$ = -0.17, p = .567) (see Table 3). In Experiment 2, there was clear evidence for an increase in both informativeness (M_d = -2.10, 95% CI [-Inf, -0.77], $t(149)$ = -2.61, p = .005) and decision noise (M_d = -1.17, 95% CI [-Inf, -0.56], $t(149)$ = -3.18, p = .001) during longer horizons, regardless of which of the four ways we defined difficulty. However, for Experiment 3, we observed non-significant differences between information parameters (z = -1.00, p = .160) and between temperature parameters (z = -0.83, p = .203) across horizons.

Discussion

Across three bandit experiments, two pre-registered, we find consistent evidence that people recognize how opportune a particular moment is to explore, and that they exploit that knowledge during their decisions. Sampling bandits in our task always involved some cost since we tied them to math problems, but only in some cases—when more trials remained—could participants opportunistically pounce or punt on that difficulty. In other words, when a participant wanted to sample a given bandit, they had no flexibility about

when to do so if it was the only free choice trial in the game. We found, however, that in comparison, facing several upcoming trials led participants to be more sensitive to the math trial difficulty. Apparently, knowing more trials remained made them more likely to wait on a bandit with a difficult math problem; likewise, it made them more eager to sample bandits whose costs they knew were cheap now but likely to soon rise. In contrast, we surprisingly found inconsistent evidence for the presence of two of the best documented exploration strategies, directed and random exploration.

The presence of opportunism in our findings links sequential decision-making research to the exploration literature. Many decision-making studies have documented that different parts of the brain are responsible for deciding what to choose compared to deciding when to choose it (Brass & Haggard, 2008; Zapparoli, Seghezzi, & Paulesu, 2017). If the timing is at least a partially independent process in many decisions, it stands to reason that it should also be a component of exploration decisions. We believe our findings to be the first to tap into this component. To be sure, the exploration literature has considered the related idea of anticipation. That is, people certainly think about future events, like upcoming trials, when making a decision—knowing more trials are to come boosts exploration in the now, because one will have more chances to later exploit whatever they explore now (Gureckis & Rich, 2018; Navarro, Newell, & Schulze, 2016; Sang, Todd, Goldstone, & Hills, 2020; Wilson et al., 2014). However, our results go a step further, showcasing the ability to plan a sequence of decisions in the context of exploration. Specifically, people may be uncertain about one kind of distribution—the expected rewards of each bandit—but they can leverage what they know about other distributions, like expected math difficulty, to opportunistically arrange their choices.

It is unclear whether opportunism was a deliberate strategy that participants deployed, or if it occurred more automatically. One could tease apart this question by subjecting participants to a working memory load while completing the task, a method Cogliati Dezza, Cleemans, and Alexander (2019) used to show that another strategy, directed exploration (which decreased under load), is a top-down process. The fact that the presence of these two strategies, opportunistic and directed exploration, did not consistently coincide with one another in the present study suggests that both may indeed be top-down and, by extension, in competition. Nevertheless, participants in Experiments 1 and 3 did not explore randomly either, even though Cogliati Dezza et al. (2019) suggest random exploration is an automatic process. Participants may have simply prioritized one attribute, the relative difficulty of bandits' math problems, over how many times they had previously explored that bandit. Future work can disentangle the strategies that go into play in settings where opportunism is leverageable.

Together, the different strength of results among random, directed, and opportunistic exploration raises questions about what constitutes a strategy. On the one hand, participants may have, explicitly or implicitly, switched between exploration strategies from trial to trial. On the other hand, they may have carried out multiple strategies concurrently. If so, they may have applied random, directed, and opportunistic exploration strategies separately to the choice options, generating a weighted score for the options. Alternatively, they may have somehow combined strategies. Work by Wilson, Wang, Sadeghiyeh, and Cohen (2020 preprint) contends that random and directed exploration are ends on a continuum of a singled shared strategy known as "Deep Exploration."; in that

model, people simulate which option to pick, and the more simulations they run, the more directed their exploration becomes. Future work could investigate whether a different holistic model could incorporate opportunistic exploration among other exploration strategies.

We should note a number of conceptual limitations in our experiment. We used large samples to obtain relatively modest effective sizes, though we note the task was performed online and we could not guarantee that participants bought into the math difficulty aspect of the task instead of cheating with an online calculator. The size of the effects also may in part be due to participants' inability to notice or keep in mind the opportunistic nature of the task, rather than deliberately ignoring whether upcoming trials would be easier or harder. If participants did approach each problem in a sort of silo, that might say more about how participants even recognize and remember opportunities and patterns rather than a reluctance to take advantage of them (Marković, Goschke, and Kiebel, 2021). Additionally, we defined opportunities in a specific way, with regard to math difficulty, and that is only one such measure of an opportunity. We alternatively could have defined opportunities in relation to availability, for example. A future study could eliminate one of two bandits on a predictable set of bandit pulls, thereby potentially encouraging a participant to pull and sample that bandit before it temporarily disappears.

Stepping back, though, our results provide compelling evidence for people's capacity and willingness to seize opportunities to explore choice options when they are relatively cheap. This strategy is perhaps the strongest affirmative answer yet to a question posed by Schulz and Gershman (2019), among others, about whether people plan exploratory behavior, as opposed to it emerging more as a heuristic. Likewise, just as opportunism is revealing of exploration, so too is exploration revealing about opportunism. It is hard to imagine a situation riper for prioritizing the "when" as exploration; if one is pre-committed to sampling several options anyways, and therefore at least to some extent indifferent between them, they will be left to focus their attention elsewhere, such as on opportune timing. It is telling that this paper behaviorally replicates a conceptually similar paradigm recently proposed in computer science (Wu, Guo, & Liu, 2018), signaling a convergence of opportunistic exploration research in disparate domains. Regardless, our participants' ability to recognize cheap and expensive moments to sample and to pounce or punt, accordingly, speaks to human sophistication in flexibly managing uncertainty. That probably bodes well for young prospects' playing time during blowouts.

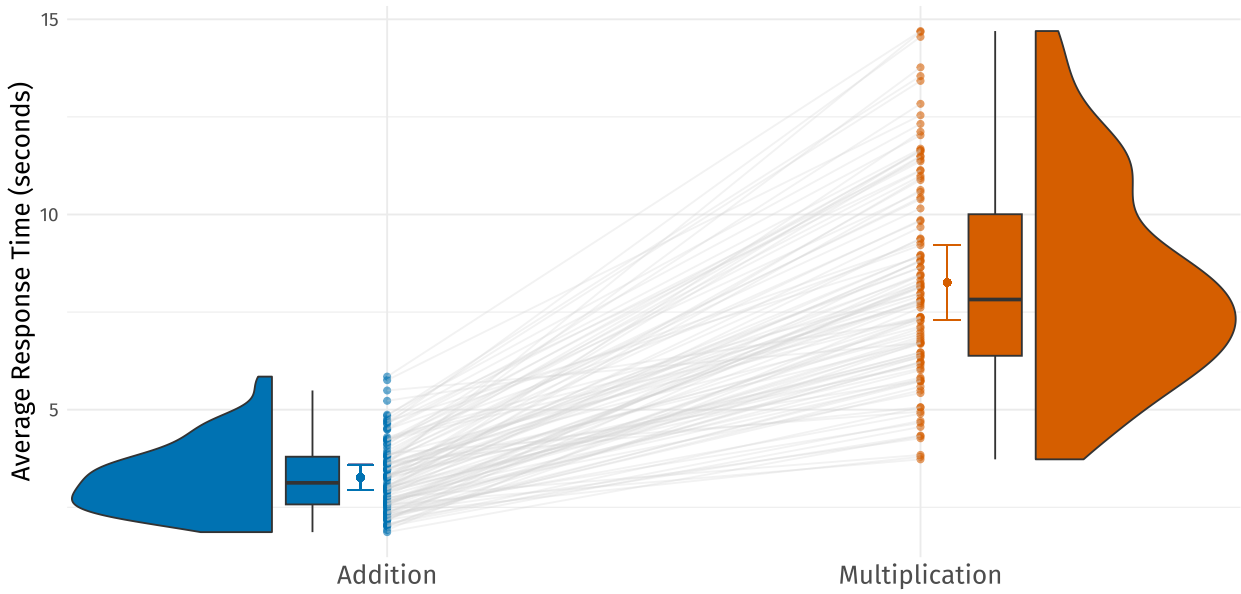

Figure 1. Individual difference in response time for correct answers to addition and multiplication problems

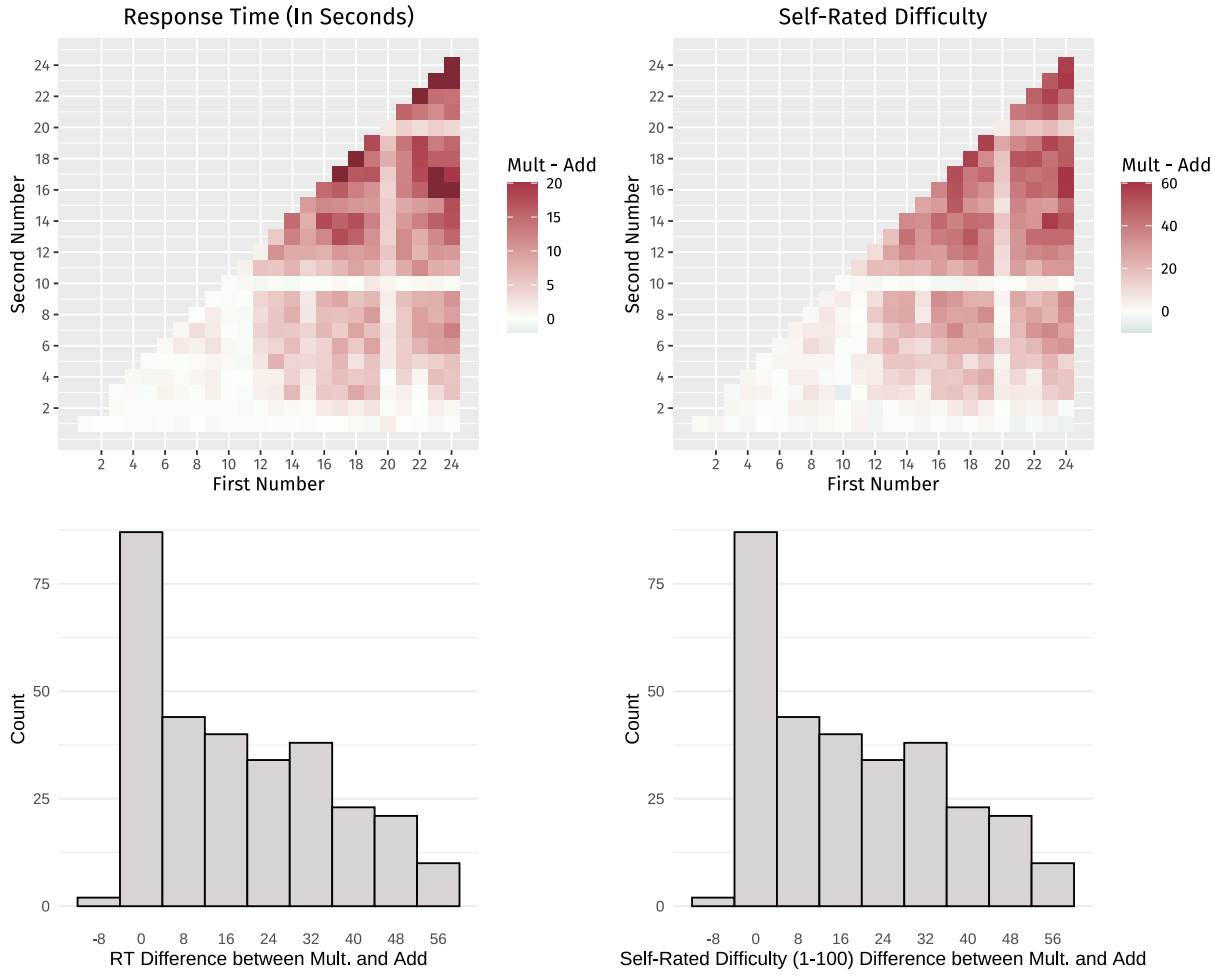

Figure 2. Difference in difficulty between multiplication and addition problems by number pairs

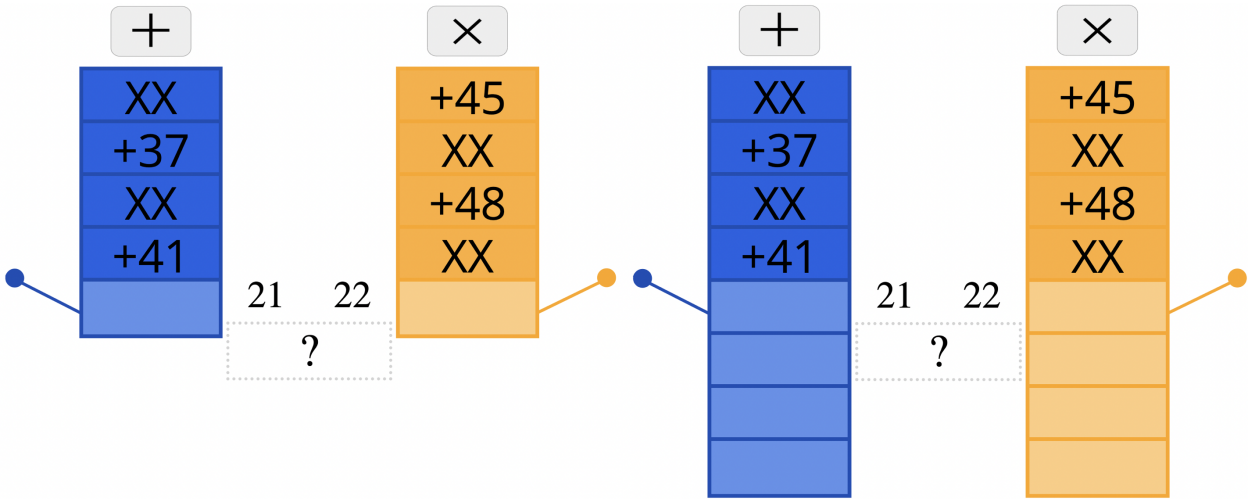

Figure 3. Example games from the experiment; short horizon on the left long horizon on right

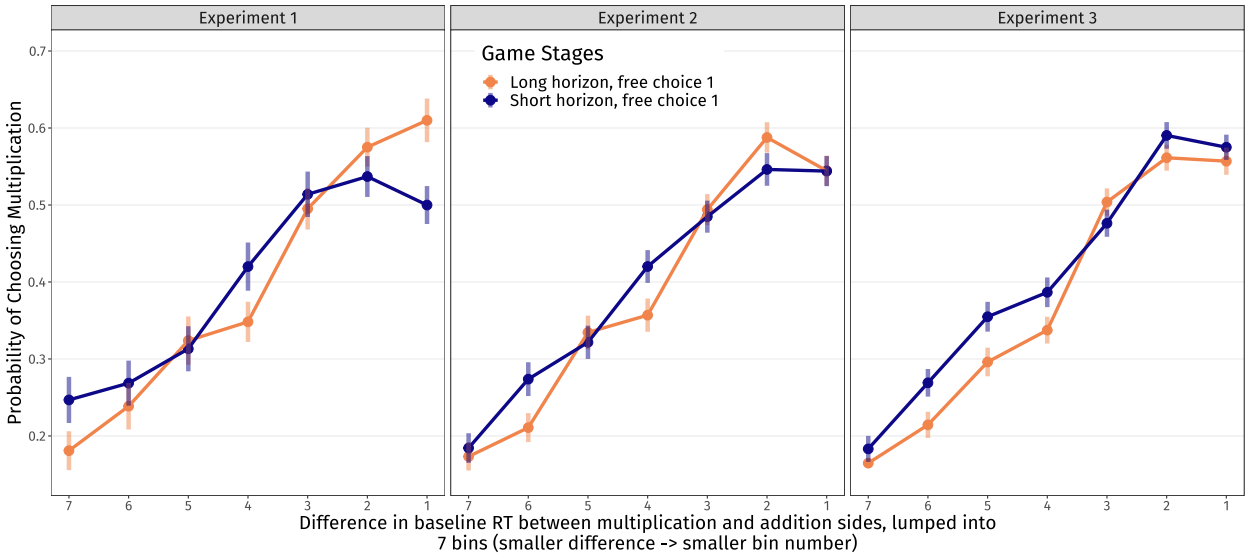

Figure 4. Facing several upcoming trials led participants to be more sensitive to the math trial difficulty

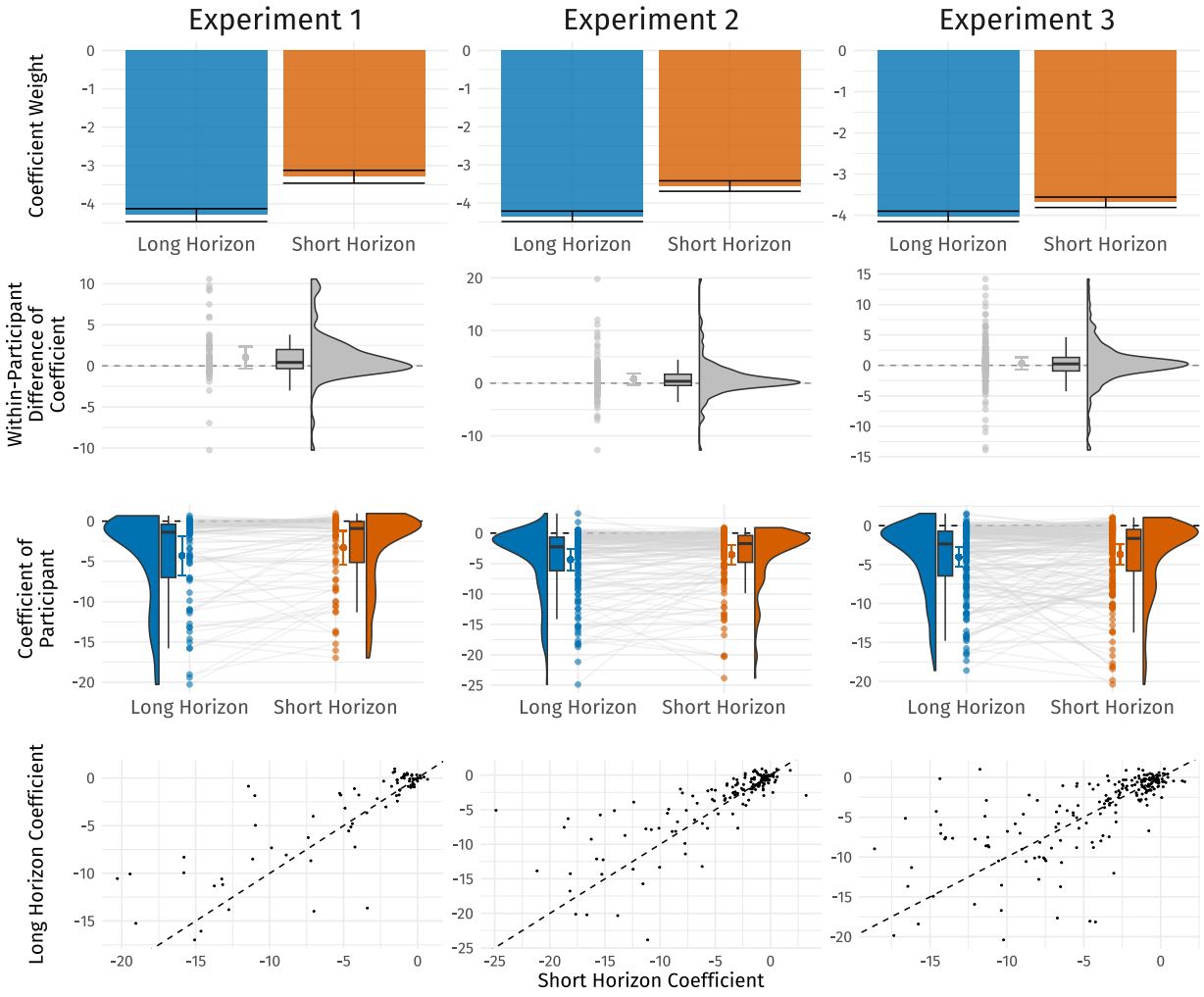

Figure 5. Estimates of the difficulty parameter, D, across three experiments

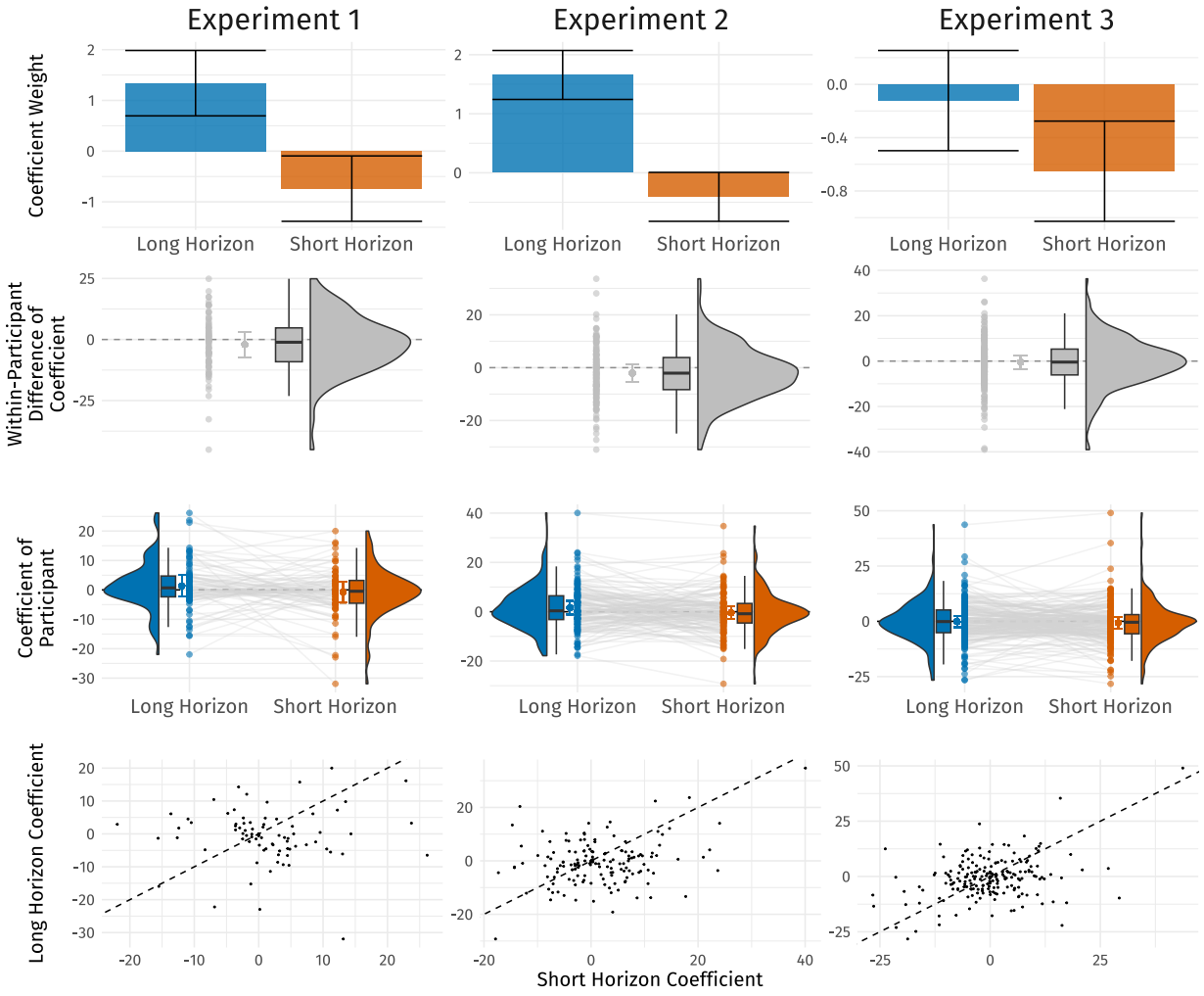

Figure 6. Estimates of the information-seeking parameter, α, across three experiments

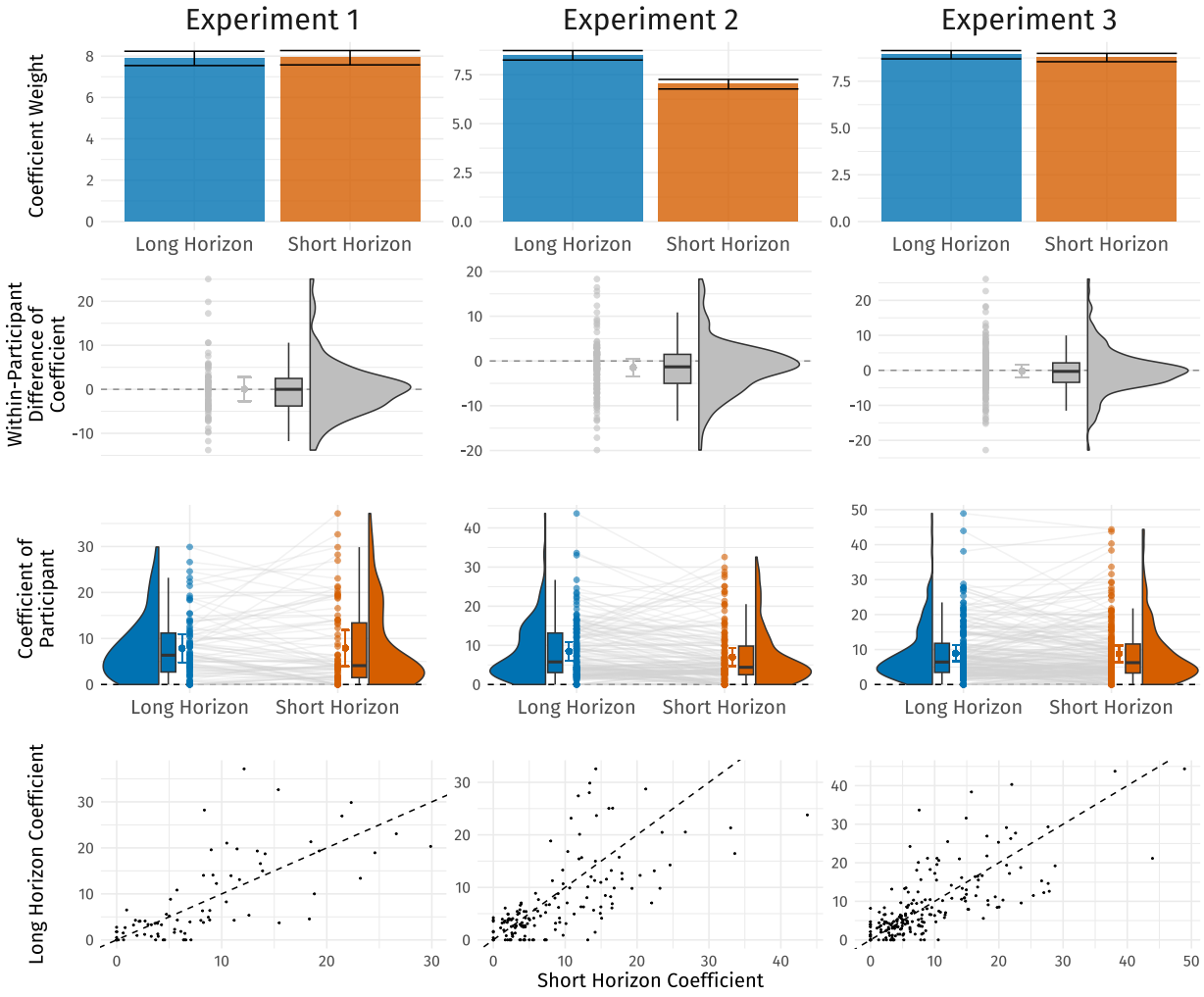

Figure 7. Estimates of the decision noise parameter, σ, across three experiments

Table 1 Estimates of the difficulty parameter, *D*, across the three bandit experiments, measuring whether sensitivity to difficulty is greater at the start of a long horizon compared to a short horizon.

Experiment (and type of pair-wise test)	Proxy for Difficulty in the Model	Pre-registered?	Parameter Estimate	95% Confidence Interval	*t* Statistic	Significance
Experiment 1 (pair-wise *t* test)	RT (Rank-ordered)		$M_d = 5.64$	[3.92, Inf]	$t(80) = 5.45$	$p < .001$
	RT (Raw)		$M_d = 1.00$	[0.44, Inf]	$t(80) = 2.97$	$p = .002$
	Self-Rated Difficulty (Rank-ordered)		$M_d = 4.79$	[3.00, Inf]	$t(80) = 4.46$	$p < .001$
	Self-Rated Difficulty (Raw)		$M_d = 0.29$	[0.09, Inf]	$t(80) = 2.37$	$p = .010$
Experiment 2 (pair-wise *t* test)	RT (Rank-ordered)	Yes	$M_d = 1.42$	[-.04, Inf]	$t(149) = 1.61$	$p = .055$
	RT (Raw)		$M_d = 0.79$	[0.34, Inf]	$t(149) = 2.89$	$p = .002$
	Self-Rated Difficulty (Rank-ordered)		$M_d = 2.15$	[0.71, Inf]	$t(149) = 2.47$	$p = .007$
	Self-Rated Difficulty (Raw)		$M_d = 0.31$	[0.13, Inf]	$t(149) = 2.81$	$p = .003$
Experiment 3 (Wilcoxon signed-rank test)	RT (Rank-ordered)	Yes	$z = 3.037$			$p = .001$
	RT (Raw)	Yes	$z = 1.650$			$p = .0495$
	Self-Rated Difficulty (Rank-ordered)	Yes	$z = 3.004$			$p = .001$
	Self-Rated Difficulty (Raw)	Yes	$z = 1.710$			$p = .044$

Table 2 Estimates of the information-seeking parameter, $\propto$, across the three bandit experiments, measuring whether participants sought out the more informative (less sampled) bandit more often at the start of a long horizon compared to a short horizon.

Experiment (and type of pair-wise test)	Proxy for Difficulty in the Model	Pre-registered?	Parameter Estimate	95% Confidence Interval	*t* Statistic	Significance
Experiment 1 (pair-wise *t* test)	RT (Rank-ordered)		M_d = -2.19	[-Inf, -0.26]	$t(80)$ = -1.88	p = .032
	RT (Raw)		M_d = -2.08	[-Inf, 0.07]	$t(80)$ = -1.61	p = .055
	Self-Rated Difficulty (Rank-ordered)		M_d = -1.99	[-Inf, -0.07]	$t(80)$ = -1.73	p = .044
	Self-Rated Difficulty (Raw)		M_d = -2.05	[-Inf, 0.01]	$t(80)$ = -1.66	p = .051
Experiment 2 (pair-wise *t* test)	RT (Rank-ordered)	Yes	M_d = -2.10	[-Inf, -0.77]	$t(149)$ = -2.61	p = .005
	RT (Raw)		M_d = -2.07	[-Inf, -0.69]	$t(149)$ = -2.49	p = .007
	Self-Rated Difficulty (Rank-ordered)		M_d = -2.20	[-Inf, -0.85]	$t(149)$ = -2.70	p = .004
	Self-Rated Difficulty (Raw)		M_d = -2.10	[-Inf, -0.73]	$t(149)$ = -2.54	p = .006
Experiment 3 (Wilcoxon signed-rank test)	RT (Rank-ordered)	Yes	z = -0.996			p = .160
	RT (Raw)	Yes	z = -0.442			p = .329
	Self-Rated Difficulty (Rank-ordered)	Yes	z = -0.843			p = .200
	Self-Rated Difficulty (Raw)	Yes	z = -0.511			p = .305

Table 3 Estimates of the decision noise parameter, σ, across the three bandit experiments, measuring whether participants explored more randomly with respect to the other parameters at the start of a long horizon compared to a short horizon.

Experiment (and type of pair-wise test)	Proxy for Difficulty in the Model	Pre-registered?	Parameter Estimate	95% Confidence Interval	*t* Statistic	Significance
Experiment 1 (pair-wise *t* test)	RT (Rank-ordered)		M_d = 0.10	[-Inf, 1.08]	$t(80)$ = -0.17	p = .567
	RT (Raw)		M_d = 0.03	[-Inf, 1.19]	$t(80)$ = 0.05	p = .519
	Self-Rated Difficulty (Rank-ordered)		M_d = -0.28	[-Inf, 0.71]	$t(80)$ = -0.47	p = .319
	Self-Rated Difficulty (Raw)		M_d = -0.33	[-Inf, 0.86]	$t(80)$ = -0.47	p = .321
Experiment 2 (pair-wise *t* test)	RT (Rank-ordered)	Yes	M_d = -1.17	[-Inf, -0.56]	$t(149)$ = -3.18	p = .001
	RT (Raw)		M_d = -1.48	[-Inf, -0.67]	$t(149)$ = -3.03	p = .001
	Self-Rated Difficulty (Rank-ordered)		M_d = -1.07	[-Inf, -0.43]	$t(149)$ = -2.78	p = .003
	Self-Rated Difficulty (Raw)		M_d = -1.45	[-Inf, -0.57]	$t(149)$ = -2.73	p = .004
Experiment 3 (Wilcoxon signed-rank test)	RT (Rank-ordered)	Yes	z = -0.830			p = .203
	RT (Raw)	Yes	z = 1.063			p = .144
	Self-Rated Difficulty (Rank-ordered)	Yes	z = -0.975			p = .165
	Self-Rated Difficulty (Raw)	Yes	z = -1.182			p = .119

Table 4 Estimates of the side bias parameter, *B*, across the three bandit experiments, measuring whether participants paid more attention to the side of the screen that a bandit was on (left or right) at the start of a long horizon compared to a short horizon.

Experiment (and type of pair-wise test)	Proxy for Difficulty in the Model	Pre-registered?	Parameter Estimate	95% Confidence Interval	*t* Statistic	Significance
Experiment 1 (pair-wise *t* test)	RT (Rank-ordered)		M_d = -0.65	[-Inf, 2.10]	$t(80) = 0.74$	$p = .770$
	RT (Raw)		M_d = -2.75	[-Inf, -0.92]	$t(80) = -2.51$	$p = .007$
	Self-Rated Difficulty (Rank-ordered)		M_d = 0.41	[-Inf, 1.84]	$t(80) = 0.47$	$p = .680$
	Self-Rated Difficulty (Raw)		M_d = -2.77	[-Inf, -0.89]	$t(80) = -2.46$	$p = .008$
Experiment 2 (pair-wise *t* test)	RT (Rank-ordered)		M_d = -1.45	[-Inf, -0.57]	$t(149) = -2.73$	$p = .004$
	RT (Raw)		M_d = -0.90	[-Inf, 0.41]	$t(149) = -1.13$	$p = .129$
	Self-Rated Difficulty (Rank-ordered)		M_d = 0.94	[-Inf, 1.86]	$t(149) = 1.67$	$p = .952$
	Self-Rated Difficulty (Raw)		M_d = -0.86	[-Inf, 0.48]	$t(149) = -1.06$	$p = .145$
Experiment 3 (Wilcoxon signed-rank test)	RT (Rank-ordered)		z = 0.012			$p = 0.505$
	RT (Raw)		z = 0.393			$p = 0.653$
	Self-Rated Difficulty (Rank-ordered)		z = 0.176			$p = 0.570$
	Self-Rated Difficulty (Raw)		z = 0.337			$p = 0.632$

References

Brändle, F., Stocks, L. J., Tenenbaum, J., Gershman, S. J., & Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*. doi: 10.1038/s41562-023-01661-2

Brass, M., & Haggard, P. (2008). The what, when, whether model of intentional action. *The Neuroscientist*, *14*(4), 319–325. doi:10.1177/1073858408317417

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi:10.3758/s13428-014-0458-y

Cogliati Dezza, I.., Cleeremans, A., & Alexander, W. (2019). Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, *148*(6), 977–993. doi:10.1037/xge0000546

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2017). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological Research*, *83*(5), 1033–1056. doi:10.1007/s00426-017-0943-x

Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore-exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, *11*(1). doi:10.1038/s41598-021-82530-8

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. doi:10.1016/j.cognition.2017.12.014

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science*, *3*(4), 275–310. doi:10.1207/s15516709cog0304_1

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. doi:10.1126/science.143.3611.1190

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience. *Psychological Science*, *21*(12), 1787–1792. doi:10.1177/0956797610387443

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263. doi:10.2307/1914185

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665–682. doi:10.1037/a0020198

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior; the Hixon Symposium* (pp. 112–146). Wiley.

Marković, D., Goschke, T., & Kiebel, S. J. (2020). Meta-control of the exploration-exploitation

dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(3), 509–533. doi:10.3758/s13415-020-00837-x

Meyers, E. A., & Koehler, D. J. (2020). Individual differences in exploring versus exploiting and links to delay discounting. *Journal of Behavioral Decision Making*, *34*(4), 515–528. doi:10.1002/bdm.2226

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. United Kingdom: Holt, Rinehart and Winston.

Navarro, D. J., Tran, P., & Baz, N. (2018). Aversion to option loss in a restless bandit task. *Computational Brain & Behavior*, *1*(2), 151–164. doi:10.1007/s42113-018-0010-8

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. doi:10.1016/j.cogpsych.2016.01.001

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Retrieved from R Foundation for Statistical Computing website: https://www.R-project.org/

Rich, A. S., & Gureckis, T. M. (2018). Exploratory choice reflects the future value of information. *Decision*, *5*(3), 177–192. doi:10.1037/dec0000074

Richardson, N., Cook, I., Crane, N., Keane, J., François, R., Ooms, J., & Apache Arrow. (2021). *arrow: Integration to Apache 'Arrow'*. Retrieved from https://CRAN.R-project.org/package=arrow

Sang, K., Todd, P. M., Goldstone, R. L., & Hills, T. T. (2020). Simple threshold rules solve explore/exploit trade-offs in a resource accumulation search task. *Cognitive Science*, *44*(2). doi:10.1111/cogs.12817

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. doi:10.1016/j.conb.2018.11.003

Schulz, E., Klenske, E. D., Bramley, N. R. & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. doi:10.1101/110486

Storn, R. (2018). *RcppDE: Global Optimization by Differential Evolution in C++*. Retrieved from https://CRAN.R-project.org/package=RcppDE

Tiedemann, F. (2020). *gghalves: Compose Half-Half Plots Using Your Favourite Geoms*. Retrieved from https://github.com/erocoar/gghalves

Westbrook, A., Kester, D., & Braver, T. S. (2013). What Is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS ONE*, *8*(7), e68210. doi:10.1371/journal.pone.0068210

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2*. Retrieved from https://wilkelab.org/cowplot/

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. doi:10.1016/j.cobeha.2020.10.001

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. doi:10.1037/a0038199

Wilson, R. C., Wang, S., Sadeghiyeh, H., & Cohen, J. D. (2020). Deep exploration as a unifying account of explore-exploit behavior. doi:10.31234/osf.io/uj85c

Wu, H., Guo, X. & Liu, X. (2018). Adaptive exploration-exploitation tradeoff for opportunistic bandits. *Proceedings of the 35th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, *80*, 5306-5314. Available from https://proceedings.mlr.press/v80/wu18b.html

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *ELife*, *6*. doi:10.7554/elife.27430

Zapparoli, L., Seghezzi, S., & Paulesu, E. (2017). The what, the when, and the whether of intentional action in the brain: A meta-analytical review. *Frontiers in Human Neuroscience*, *11*. doi:10.3389/fnhum.2017.00238