Skilled Bandits: Learning to Choose in a Reactive World

Jared M. Hotaling (jared.hotaling@gmail.com), Danielle J. Navarro, and Ben R. Newell
School of Psychology, University of New South Wales
Sydney 2052, Australia

Abstract

In uncertain environments we must balance our need to gather information with our desire to exploit current knowledge. This is further complicated in reactive environments where actions produce long-lasting change. In three experiments, we investigate how people learn to make effective decisions from experience in a dynamic four-armed bandit task. In contrast to the diminishing rewards found in most previous studies, options were framed as skills that developed greater rewards when chosen. We find that most individuals learn effective strategies for coping with reactive environments. We present a psychological model positing that decision makers move through three distinct processing phases, and show that it accounts for key behavioral patterns across experiments.

Keywords: decision making; dynamic environments; explore-exploit dilemma; decisions from experience

Introduction

Many important choices in life lack detailed descriptions and require that decision makers learn about alternatives through experience. In these situations, one must balance the need to gather information through exploration, with the desire to exploit current knowledge (see Mehlhorn et al., 2015). Consider the dilemma of a student choosing which classes to take. “Commercial Law” and “Introduction to Programming” both sound interesting and potentially lucrative over time, so how should one invest in one’s long term future? Careers in law or IT could both be rewarding, but it might depend on one’s aptitude, so it is wise to consider how one’s skills are likely to develop with time.

This scenario presents a difficult decision making problem: the student needs to estimate her long run rewards, based on how her skills are likely to develop over time (in programming or in law), using only information about how well she performs at the very beginning of the learning sequence. Unfortunately, some skills can be acquired quickly and can be rewarding after very little investment, whereas others mature slowly and provide a delayed reward. It may be difficult for decision makers to discern that the eventual reward from a slow-maturing option is greater.

This problem provides an example of a *reactive* environment. Not only do the rewards associated with an option change over time (e.g., salaries fluctuate), they depend on one’s choices (e.g., how much effort one spends learning the intricacies of corporate law). There are many examples of this kind of reactivity: the profitability of a product is shaped by the way it is advertised, the output of a farm increases as cultivation techniques are optimized, and so on. Critically, in a reactive environment our actions produce long-lasting changes in the world. In such an environment it is important to learn what options exist and how one’s actions influence these options over time.

Previous research on decisions from experience has tended to focus on static environments. Here the optimal strategy is to “front-load” one’s exploration to maximize expected long-term rewards (Tversky & Edwards, 1966). This is not the case when payoffs change over time. In these dynamic environments, one must monitor the evolution of each option in case changes occur (e.g. Navarro, Newell, & Schulze, 2016). Even research involving dynamic environments (e.g. Gureckis & Love, 2009) has tended to focus on exploitation-dependent *diminishing* rewards. In these tasks, the rewards associated with an option decrease each time it is chosen.

In the student dilemma described above the opposite pattern holds: rewards increase each time an option is chosen. In this paper, we focus on sequential decision making in a reactive environment that is inspired by this scenario. Decision makers must strategize their choices so that they quickly find the options with the most potential for growth. With enough experience in the same environment, they can learn to identify promising options more quickly based on previously encountered growth trajectories, allowing more time for exploitation and growth.

The Skilled Bandit Task

Our experimental task presented participants with several options with unknown reward distributions that must be learned from a sequence of consequential choices. Although there are studies that employ bandit style problems in a “resource depletion” scenario, where the rewards decrease with choice (Gureckis & Love, 2009) or the availability of options changes over time (Ejova, Navarro, & Perfors, 2009; Shin & Ariely, 2004), comparatively little is known about how people solve a “resource cultivation” problem where rewards increase as a function of repeated usage.

We framed our study as a simulation of a “skill development” scenario in which participants repeatedly chose which of four skills to develop. Each time a skill was chosen it produced a reward. However, with each action the chosen skill would incrementally develop so that it would offer a higher reward in the future. Options developed differently, with some initially offering low rewards but growing substantially. Participants learned about the options by sampling (i.e. choosing) each and observing the payoffs they received. Since the number of choices in each game was fixed, efficient sampling was important. Participants completed five games in the same environment. Importantly, in Games 2-5 participants had prior knowledge of what payoff functions existed in the environment. Combining this prior information with the payoffs observed in the current game allows for better inferences about the long term value of each option.

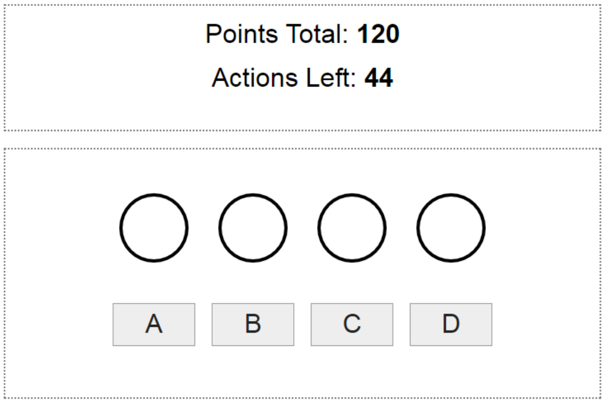
Below, we present findings from three experiments using the skilled bandit task, along with a cognitive model.

Experiment 1

The choice environment involved four options defined by their growth trajectory: low slow (LS), low fast (LF), high slow (HS), and high fast (HF). We expected that some individuals would be attracted by the initial gains of LF, but would eventually learn to choose the optimal HF. Half of participants received an additional warning that some options may seems good at first, but fail to develop. We predicted that this warning would improve performance.

Method

Participants 201 US-based participants (85 female) were recruited via Amazon Mechanical Turk. The average reported age was 35.35 years (SD = 11.56). The experiment lasted 10 minutes, and participant received $1.70.



Points Total: **120**

Actions Left: **44**

Figure 1: The choice screen in the skilled bandit task after making six choices.

Procedure & Materials Participants completed the experiment online. They were shown a choice screen from the game (Figure 1), and were told to choose one of four options labelled A to D. After clicking on a button, participants received a payoff of between 1 and 99 points. The amount was displayed inside the clicked button for 800 ms. Participants were told that the goal of the game was to "win as many points as possible", and that they had a total of fifty actions (clicks) each game. The points total and number of remaining actions were displayed on screen at all times. Half of participants received only *vague* instructions that "some ways of playing will work better than others".

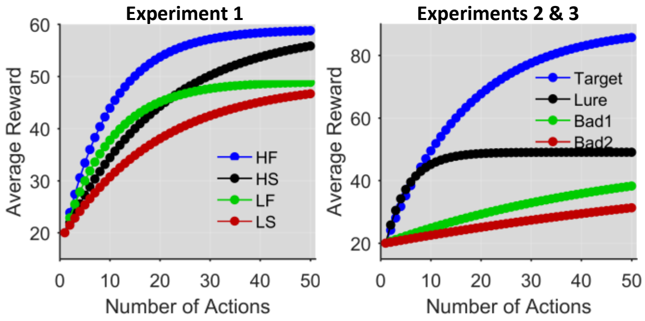


Figure 2: Payoff functions for Experiments 1-3.

The remaining 50% of participants in the *detailed* instruction condition were also told:

> *Each time you practice a skill (i.e. click on a button) you get better at it (i.e. win more points). But the rate at which you improve might differ across the different skills. Just like in life, some skills are quick to learn - you improve rapidly - but in the end they may not be as rewarding as other skills which are slower to learn but, ultimately, deliver larger rewards. To win the most points you need to find out which skill (button) will provide the best rewards overall.*

Feedback was given at the end of each game, indicating the total points earned, as well as the maximum (2600 points) and minimum (1400 points) possible scores. At the beginning of each game options were randomly assigned to buttons A-D. At the end of the experiment, participant were given a completion code to claim payment via Amazon.

The expected payoff of each option was determined by the number of times it was chosen. Figure 2 shows the payoff functions for the four options. To make the differences between options less obvious, Gaussian random noise (M = 0, SD = 3) was added to the expected payoffs.

Results

Figure 3 shows the number of points earned across games and conditions[1]. To our surprise, participants in the detailed condition earned only slightly more points per game (M = 2,055.39, SD = 11.73) than those in the vague condition, (M = 2,052.30, SD = 11.59), t(198) = 0.07, p = 0.53. A chi-square test confirmed that mean choice proportions were not significantly different across conditions, X^2(3, N = 50,250) = 3.45, p = 0.33, so we collapse across instructions conditions for all subsequent analyses. The top row of Figure 4 shows that participants began each game by choosing from each option but soon developed a preference for HF. It is noteworthy that the LF option shows a smaller decrease across trials than HS and LS, suggesting that some participants were drawn to its rapid early growth.

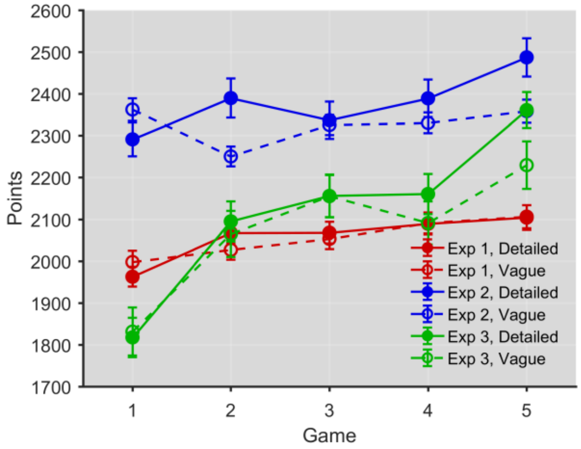


Figure 3: Mean points earned in each game and experiment. Error bars indicate standard errors.

[1] We excluded choices from games where the same option was chosen on every trial. These did not substantially affect our results.

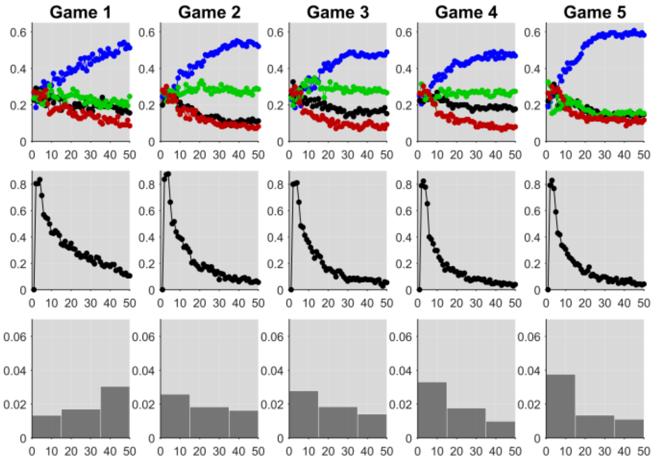


Figure 4: Results from Experiment 1, across trials (x-axis) and games (columns). The top row shows mean choice proportions. The middle row shows switching probability. The bottom row shows stopping point distributions.

To better understand the strategies people use to solve the skilled bandit task, we also analyzed people's exploration and exploitation behavior. The middle row of Figure 4 shows the probability of *switching*, defined as choosing a different option from the one chosen on the previous trial. Clearly participants began each game with a large degree of exploration, switching between options often. However, this exploration quickly gave way to more exploitative behavior, characterized by lower switching rates. Participants also engaged in less switching across games, suggesting that their exploration became more efficient with experience because they required fewer samples from each option. To test the effects of game and trial on switching we used a multilevel generalized linear mixed-effects model with a binomial distribution and logit link function. In the model *game* and *trial* were fixed effects, with random effects at the subject level for both predictors. We found significant effects of both trial (β = -0.155, *SE* = .008, p < .001) and game (β = -0.370, *SE* = .044, p < .001). A similar pattern can be seen if one considers the point at which an individual stops exploring and chooses the same option for all remaining trials. The bottom row of Figure 4 shows the distribution of these stopping points, defined by the trial containing the final switch of the game. In Game 1 this distribution was negatively skewed, with many participants stopping late (*M* = 30.14, *SD* = 14.80). However, across games, stopping points shifted earlier, with a mean stopping point of 17.42 (*SD* = 13.91) in Game 5. A repeated measures analysis of variance (ANOVA) confirmed that stopping point decreased across games, *F*(4,608) = 33.23, *p* < .001.

Discussion

Experiment 1 show that people are capable of learning to make good decisions in a dynamic risky choice environment characterized by resource growth. Within the first game the majority of participants quickly found the option offering the highest payoffs. More impressively, across games, participants improved their performance. Earning more points with less exploration, participants demonstrated that they accumulated knowledge about the dynamics of the environment across games, reducing their need to explore.

The shifting stopping point distributions shown in Figure 4 points to a gradual increase in front-loaded exploration strategies suggesting that participants focused their efforts on distinguishing the options from one another at the beginning of a game. Toward the end of a game, participants increasingly stopped exploring and committed to exploiting the best option. Such front-loaded exploration may also explain why LF was the second most chosen option. LF's payoffs increased rapidly at the beginning of each game, so participants who stopped exploring early on would likely form an erroneously positive impression of this option relative to the others.

Experiments 2 and 3

In Experiments 2 and 3, a *target* option, T, offered the highest total payoffs but was virtually indistinguishable from a *lure* option, L, early on. This option also offered fast increasing payoffs for the first ten actions, after which its payoffs plateaued. To compensate for the increased task difficulty produced by the lure, the remaining two bad options, B1 and B2, were made clearly inferior. These changes allow a strategy where decision makers first explore their options in order to identify and cull the clearly inferior alternatives, then try to distinguish the target from the lure, and exploit it for the remaining trials.

The environmental dynamics in Experiment 2 matched those from Experiment 1: each time an option was chosen its expected payoff incremented according to its payoff function, otherwise options remained unchanged. In Experiment 3 we explore a different type of environment, where unchosen options decrement. Inspired by the "use it or lose it" nature of many real world skills, we incorporated a simple forgetting function in which each time a participant failed to choose an option the expected payoff decreased according to its payoff function. Thus, focusing one's efforts on one skill meant becoming "rusty" at other, neglected skills. We expected this additional complication to increase the difficulty of the task, especially in Game 1. These dynamics also complicate the mapping between actions and outcomes, and require that individuals update option values in memory. This poses a challenge to many simple reinforcement learning (RL) models that rely on immediate feedback and lack robust internal representations. We expected participants in Experiment 3 to eventually adapt to their environment by switching between options less so as to minimize decrements.

Method

Participants 198 and 300 participants (69 and 121 female) were recruited via Mechanical Turk for Experiments 2 and 3. The average age for participants was 34.98 years (*SD* = 11.15) and 36.40 (*SD* = 11.88). Experiments 2 and 3 each lasted 10 minutes and participant received $1.70.

Procedure & Materials The experimental task, stimuli, and instructions were identical to those of Experiment 1.

Approximately half of participants were randomly selected to receive either vague or detailed instructions. Figure 2 shows the expected payoff functions for the four options in Experiments 2 and 3. In Experiment 2, the expected payoff of each option was determined by the number of times it had been chosen, just as in Experiment 1. In Experiment 3, each time an option was chosen, its expected payoff incremented, but the expected payoff for each unchosen option decremented one action. Options were not allowed to decrement below their initial values. Random noise was added to the expected payoffs as in Experiment 1.

Results

Figure 3 shows that for Experiment 2 participants in the detailed instructions condition earned more points than those in the vague instructions condition, $t(196) = 1.82$, $p < 0.05$ (one-tailed). The same pattern appeared in Experiment 3, but was not significant, $t(298) = 1.24$, $p = 0.11$. Participants earned more points in Experiment 2 (M = 2,352.74, SD = 14.41) compared to Experiment 3 (M = 2,089.47, SD = 16.70), $t(496) = 9.07$, $p < 0.001$ (two-tailed). In both cases detailed instructions increased choices for the target and decreased choices for the lure.

Figures 5 and 6 show the major experimental results collapsed across instructions conditions. Participants developed a strong preference for T and L, though they chose these more in Experiment 2 ($M = 0.85$, $SD = 0.13$) than Experiment 3 ($M = 0.77$, $SD = 0.23$), $t(496) = 7.20$, $p < .001$ (two-tailed). This difference was driven by performance in early games. In Experiment 2, the probability of choosing T or L increased from 0.84 to 0.87 across games, $F(4,664) = 1.11$, $p = 0.35$. In Experiment 3, this increase was more striking; rising from 0.70 to 0.83, $F(4,716) = 7.64$, $p < .001$.

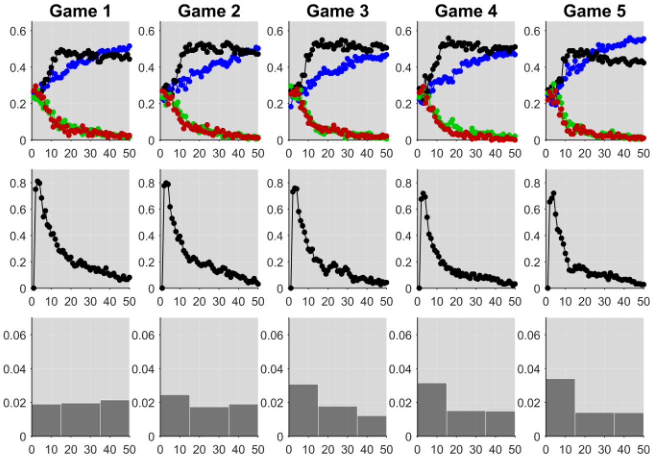


Figure 5: Results from Experiment 2.

Exploration was similar across experiments, though Experiment 3 saw larger changes across games. Figures 5 and 6 show that the probability of switching decreased across trials. Across games, sampling became more efficient and people switched less. This was more notable in Experiment 3, where the probability of switching in Game 5 was 0.12 (SD=0.15) compared to 0.17 (SD=0.18) in Game 5 of Experiment 2. This pattern repeats for stopping point, with both experiments showing a shift to earlier stopping, but with larger changes in Experiment 3. Mean stopping point decreased from 26.20 (SD=14.58) to 19.33 (SD=14.15) in Experiment 2, and from 30.35 (SD=14.91) to 17.45 (SD=13.35) in Experiment 3. Repeated measures ANOVAs confirmed that participants stopped earlier across games in Experiment 2, $F(4,664) = 11.39$, $p < .001$, and Experiment 3, $F(4,716) = 29.14$, $p < .001$.

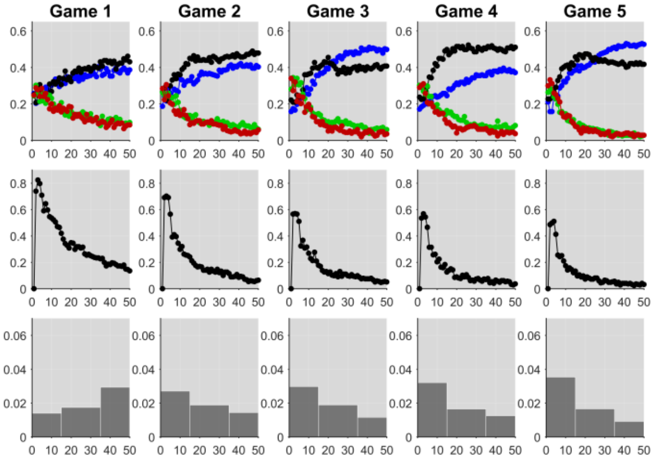


Figure 6: Results from Experiment 3.

The "use it or lose it" dynamics of Experiment 3 meant that that exploration was best done in "streaks"; selecting an option multiple times to get a sense of its current value and growth rate. Figure 7 shows that people adopt this strategy. It depicts the mean streak length calculated across a moving window of 20 trials, where a streak is defined as a sequence of consecutive choices for the same option. Comparing the experiments supports our earlier findings that participants in Experiment 3 showed more change across games, with mean streak length increasing from 7.97 (SD=5.62) to 14.06 (SD=4.55) across games compared to an increase of only 9.43 (SD=5.42) to 12.87 (SD=5.31) in Experiment 2. The effects of environmental dynamics are particularly evident at the beginning of each game. In Experiment 2 streak length across the first twenty trials increasing modestly from 2.90 (SD=2.37) to 4.48 (SD=3.62), $F(4,664) = 9.70$, $p < .001$. In Experiment 3, however, participants began Game 1 with short streaks (M=2.57, SD=2.35), but by Game 5 had learned to concentrate on one option at a time (M=6.25, SD=4.65), $F(4,716) = 28.38$, $p < .001$.

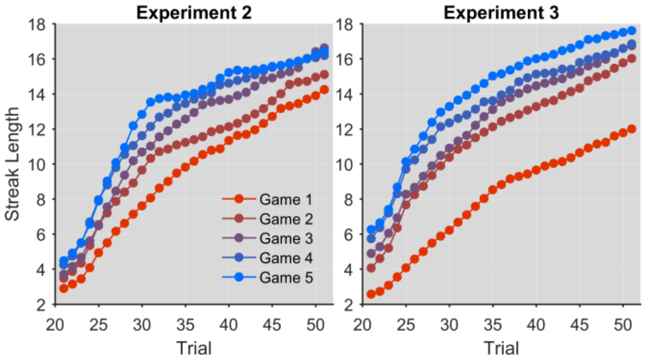


Figure 7: Mean streak length in a moving window of twenty trials.

Discussion

In Experiments 2 and 3, participants developed a strong preference for Options T and L over B1 and B2. However, they often preferred the suboptimal L most of all. Perhaps participants failed to learn that T offered better long-term rewards than L due to insufficient late-game exploration. However, another possible explanation is that participants understood the difference between T and L, but could not distinguish between them before reach a “point of no return”, after which it would be better to continue with L than switch to an “undeveloped” T.

Comparing the experiments, performance in Game 1 of Experiment 3 was adversely affected as participants struggled to understand the relationship between their choices and payoffs. Maximization (i.e. choosing T) and points totals were subsequently low in Game 1 of Experiment 3. Participants in Experiment 2 had an easier time distinguishing the options in Game 1. This would explain why more people were drawn to the lure—which offered the highest payoffs at the beginning of each game—in Experiment 2 compared to Experiment 3, where early advantage of L over T would have been harder to notice.

By Game 5, however, performance in Experiment 3 had improved substantially, with participants finding the target option and earning points at rates similar to what we observed in Experiment 2. Indeed, behavior in Experiment 3 was characterized by substantial changes across games as participants adapted to their environment. This can be seen in stopping point and streak length (especially for the first twenty trials). In Game 1, participants were more exploratory—with more switching and later stopping compared to in Experiment 2—but by Game 5 they had prioritized exploitation of their accumulated knowledge, and showed less exploration compared to in Experiment 2. Thus, participants eventually achieved comparable performance, despite being in a less advantageous environment.

A Cognitive Model for Skilled Bandits

In contrast to many simple RL models, the *discrete state model* (DSM) does not rely on one continuous process, but rather posits that decision makers can possess three distinct processing states[2]. In the *explore state*, agents sample each option a minimum number of times. After each option has been sufficiently explored, the agent enters the *exploit state*. Here it repeatedly chooses the option that yielded the highest mean payoff during exploration. Finally, some agents then enter a *monitor state* in which they check if the option being exploited ceases to develop. If this happens, the agent switches to exploiting the option that produced the second highest mean payoff during exploration. Thus, DSM instantiates the simple intuition that participants solved the skilled bandit task by briefly surveying their options, quickly transitioning to exploiting the most promising option, and (in the case of some individuals) switching to the next best option if the first stops developing.

The model also proposes two types of processing differences across individuals. First, the minimum number of exploratory choices per option differ, with individual values drawn from a normal distribution, $N(3,1)$. Second, agents may or may not enter the monitor state. Most individuals simply exploit the best option (after exploration), while others check if it plateaus. The model represents this monitoring by fitting a line to the ten most recent payoffs produced by an option. If the slope drops below 0, the agent switches to exploiting the second best option for the remaining trials. We use ρ to denote the proportion of individuals who enter a monitor state.

One of our main objectives is to understand the learning that occurred across games. Since we hypothesized that participants became more aware of the underlying payoff distributions over time, we expected that the number of individuals monitoring for plateaus would increase across games. We therefore explore the predictions of the DSM across a range of ρ values to see how well this captures the behavioral changes we observed across games.

Although the DSM did well to capture behavior in Experiment 1, for the sake of brevity we focus here on Experiments 2 and 3. Figure 8 shows predicted choices for Experiment 2 are quite accurate. The model reproduces the crossing over of T and L for a range of ρ values. This effect is more pronounced when a greater proportion of agents are monitoring for plateaus. Predicted switching rates match those observed in the data, although with slightly more early switches and fewer late switches. This is also reflected in stopping point distributions, which show that moderate values were more frequent than was observed. These misfits could be improved by introducing additional stochastic variability within or between individuals.

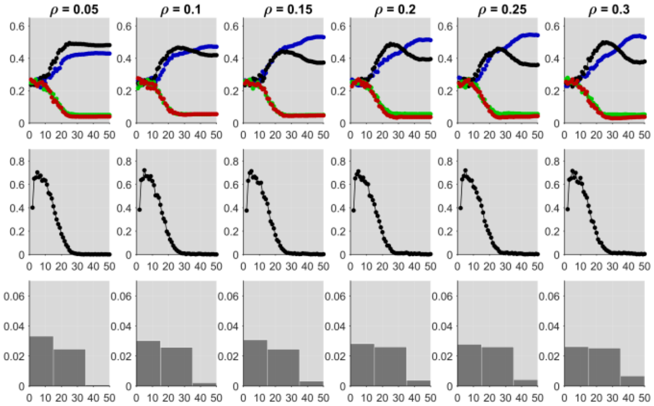


Figure 8: Predictions of the DSM in Experiment 2.

Figure 9 shows that the DSM, like many participants, found it harder to identify the advantageous options in Experiment 3. In line with our observations, preferences for T and L were weaker than in Experiment 2. That said, the model performs worse than participants, who rarely chose B1 and B2 at the end of a game. Qualitatively, predicted sampling behavior matched participants’, although the model produces more midgame switching and stopping.

[2] We also tested several simple RL models, but none performed as well as the DSM.

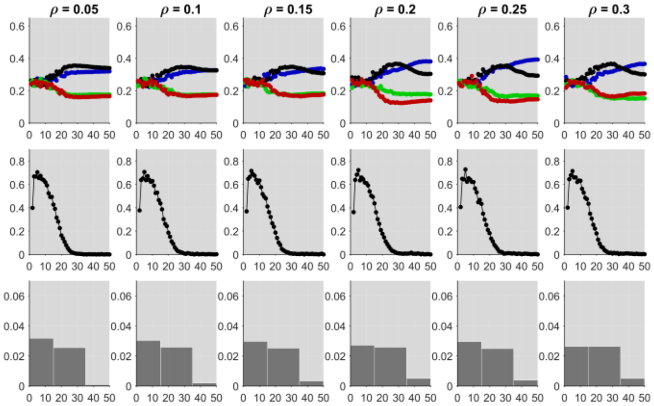


Figure 9: Predictions of the DSM in Experiment 3.

General Discussion

In three experiments, participants learned to make adaptive choices in a skilled bandit task. Performance was good, with most individuals exploiting options with high payoffs and growth. Learning across games was characterized by increased efficiency in sampling (e.g. less switching, earlier stopping). With experience, participants were able to find good options with less exploration. Impressively, this reduction in sampling coincided with greater maximization, indicating that participants had learned to make better decisions using less information. The effects of instructions were small and did not occur in all experiments suggesting that vague instructions regarding skill development, coupled with direct experience of making choices, was sufficient to understand the task dynamics. Indeed, if verbal instructions at the beginning of the experiment affected behavior, this was quickly overshadowed by the learning that occurred in the task. Experiments 2 and 3 showed that participants adapted their strategies in response to the choice environment. When payoffs obeyed a "use it or lose it" rule (Experiment 3), participants learned to minimize exploration—sampling less, stopping earlier, and choosing in streaks—in order to avoid payoff decrements. This explains why performance across experiments was so different in Game 1, but was quite similar by Game 5.

However, learning was imperfect and participants often failed to choose the option that offered the maximum payoff. In part, this stems from the inherent difficulty of the skilled bandit task – where decision makers must act on incomplete information – though it also points to the use of simple strategies for learning and making decisions.

Modeling Insights

To better understand these strategies, we developed the DSM based on the idea that decision makers complete each game in three stages. The model generally reproduced the observed choice patterns and sampling behavior. However, it failed to fully capture behavioral changes across games, indicating that additional learning mechanisms are needed to explain how processing became more efficient (i.e. more accurate with less exploration). That said, the success of the DMS – with its focus on discrete cognitive processes – presents a challenge to RL models that rely on a single, continuous learning process.

Conclusion

Our findings suggest that, with experience, individuals can learn to thrive in complex, uncertain, reactive environments. Across three experiments, individuals learned to make adaptive decisions. Across games, they became more efficient, earned greater rewards and with less exploration. This suggests that participants accumulated knowledge about the developmental trajectories of skills. As this knowledge increased, participants were better able to make inferences about unexplored options, allowing them to improve their performance on subsequent games and reducing their need to explore. The success of the DSM in using distinct processing stages provides a promising avenue for future investigations into people's strategies for leveraging past experience for effective decision making.

References

Ejova, A., Navarro, D., & Perfors, A. (2009). When to walk away: The effect of variability on keeping options viable. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 1258-1263.

Gureckis, T. M., & Love, B. C. (2009). Learning and noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology, 53*, 180-193.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A.,…Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*(3), 191-215.

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.

Shin, J., & Ariely, D. (2004). Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management science, 50*(5), 575-586.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology, 71*, 680.