Peeks and Keeps: A new paradigm for studying the exploration-exploitation trade-off

Nathaniel D. Phillips and Hans Joerg Neth
University of Konstanz

Daniel Navarro
University of Adelaide

Abstract

Many important decision tasks involve an exploration-exploitation trade-off, where organisms have the competing goals of gaining new information (exploration) to improve future decisions, and acting on existing information (exploitation). The most common paradigm to study this trade-off experimentally is the n-armed bandit, where decision makers reap real costs and rewards on every trial. We suggest that, unlike the n-armed bandit, many real world tasks allow decision makers to explore options (such as stock price changes) without reaping any costs or rewards. To address this, we introduce a new experimental paradigm called "Peeks and Keeps" that combines aspects of the n-armed bandit with the 'bet-observe' task (Tversky & Edwards, 1966). Unlike the n-armed bandit, Peeks and Keeps gives decision makers the option of explicitly separating exploration and exploitation behavior, where exploration provides only information but no costs or rewards, and exploitation gives both information and costs and rewards. This paradigm not only increases the empirical validity of the n-armed bandit, but also provides researchers with an explicit measure of exploration that is hidden in other paradigms.

Keywords: exploration, exploitation, decisions from experience, decisions under uncertainty

Introduction

One day in the 1980s, a man named Martin Frankel walked into a stock broker's office in Toledo Ohio and said he wanted to invest in the stock market. Frankel explained to the owner of the office that after studying the market for years, he had develoepd a unique insight into the market – to the point where he was cetain he could beat it. The owner was so impressed by Frankel's vision that he actually hired Frankel as a trader. Over the next 6 months, Frankel all-but lived at the broker's office – meticulously studying the market and developing increasingly ingenious trade ideas. However, after 6 months, the owner became increasingly frustrated with Frankel. Despite all of his 'brilliant' theories

and plans, Frankel never actually *acted* on any of his theories by making a trade. By his own admission, Frankel suffered from what he called 'trader's block' wherein he had a hard time 'pulling the trigger.' Because of his lack of action, Frankel was fired from the small broker's office [1].

Why did Frankel fail as a legitimate broker? On the one hand, patience is a virtue – one that Frankel had in spades. If Frankel had acted more impulsively and invested all of the firm's money in one 'hot' stock on his first day, he would may been fired for putting the entire firm at risk on too little knowledge. Instead, Frankel's failure as a legitimate stock broker exemplifies the cost of waiting too long to act.

This example exemplifies a common trade-off between acquiring information without risk, and acting on existing information with risk. This trade-off occurs in many areas of life.

- Mate Selection: A person must decide how long to get to know someone via dating before committing to a long-term relationship. Too little dating could lead to marring a poor match, while waiting too long could lead to no marriage at all.

- Food Foraging:

- Athletic training: How long should I train in boxing before I schedule my first real match?

Exploration-exploitation trade-off

Many of the most important real world decisions require individuals to reap consequences from several risky options that probabilistically give rewards and punishments. In many tasks, these decisions are made under uncertainty, where the probabilities and magnitudes associated with options are *a priori* unknown. In order to learn about options, organisms can engage in active search which improves the quality of their impressions of options. However, search can come at a cost, such as the missed opportunity to receive rewards from known options. For example, in trying a new restaurant, one forgoes the opportunity to have a meal at her (current) favorite restaurant.

This conflict between obtaining new information and acting on existing information is known as the exploration-exploitation trade-off. The exploration-exploitation (EE) trade-off is one of the most widely studied aspects of decision making from human to non-human organisms. The exploration-exploitation trade-off represents a goal conflict in decisions under uncertainty, where an organism is trying to maximize its long term rewards from *a priori* unknown options. On the one hand, individuals want to explore options by gaining as much information as possible to improve the quality of their future decisions. On the other hand, they want to *exploit* options by acting on existing information in order to increase short-term rewards.

One of the most widely used experimental tasks used to study the exploration-exploitation conflict is the n-armed bandit. In an n-armed bandit, participants have a fixed number of trials to select an option and experience a consequential reward.

---

[1] Frankel's story would have been forgotten if not for the fact that he went on to earn over $200 Million dollars in a series of investment scams and Ponzi schemes. He is currently serving a XX year prison sentence.

Purely epistemic versus pragmatic actions

- Neth and Müller (2008) distinguished between two types of actions, epistemic and pragmatic. Epistemic actions are those that result in information rather than punishments or rewards, while pragmatic actions are those that lead to punishments or rewards. Exploration is assumed to be an epistemic action while exploitation is a pragmatic action.

- One can easily imagine real-world cases where people explicitly engage in purely epistemic actions. For example, imagine a person who wishes to learn about the stock market prior to risking any real money. He can do this by viewing sequential returns from several stocks and observing their risk. Alternatively, a new resident to a town can learn about local restaurants by asking her neighbors about their recent experiences. In all of these cases, the actor is learning about options without reaping consequences.

- Clearly these purely epistemic actions are both psychologically and behaviorally distinct from pragmatic actions, where one obtains *both* information and immediate consequences. For example, our stock investor who starts investing his money into stocks will then not only learn about their performance, but also reap gains and suffer consequences. Similarly, the new town resident who starts frequenting local restaurants will continue learning about them but also experience immediate pragmatic outcomes.

- Somewhat puzzlingly, paradigms that have been used to study exploration-exploitation trade-off in humans has largely ignored behavioral differences in epistimic and pragmatic actions. In the N-armed bandit task, players are only allowed to engage in one type of behavior - choice, which always provides both epistemic and pragmatic rewards. Players are not given the option to engage in purely epistemic actions.

- This can lead to erroneous inferences. The same choice behavior could be interpreted as either resulting from an epistemic or pragmatic motivation. Until now, researchers have had to use computational cognitive modeling techniques to attribute choices post-hoc to either an epistemic or pragmatic underlying goal.

- We believe a new paradigm is needed. One where individuals have the option to explicitly explore or exploit options. This task will not only be a better model of many real-world decision tasks than previous paradigms, but will also allow researchers to explicitly observe behavior consistent with purely epistemic goals.

Combining three paradigms

Paradigm	EE Tradeoff	Pure Exploration	Pure Exploitation	Alternation
N-Armed Bandit	Yes	No	No	Yes
Sampling Paradigm	No	Yes	Yes	No
Bet-Observe	Yes	Yes	Yes	Yes
Peeks and Keeps	Yes	Yes	No	Yes

In a multi-armed bandit task, participants choose between multiple a priori unknown options over several trials and receive rewards (or costs) on each trial. Because decision makers reap consequences on every trial, the n-armed bandit task does not allow purely epidemic actions. The Iowa Gambling Task (IGT) is one famous example of this paradigm. Using cognitive models such as the expectancy-valence model, researchers have used the IGT to study cognitive mechanisms such as loss-aversion, recency, and choice consistency in both healthy and non-healthy individuals (Yechiam, Busemeyer, & Stout, 2005).

Two paradigms have been used to study purely epistemic actions: the sampling paradigm of decisions from experience (Hertwig, Barron, Weber, & Erev, 2004) and the bet-observe task (Tversky & Edwards, 1966). Like the n-armed bandit task, both paradigms present participants with multiple, a priori unknown options. In the sampling paradigm, participants can then sample from options, without consequence, as many times as they would like before making a single consequential choice. Here, participants engage in a self-determined number of purely epistemic actions strictly prior to a single purely pragmatic action. After making their final choice, participants receive the consequences from their choice but cannot continue to observe. Thus, in the sampling paradigm observation strictly occurs prior to exploitation with no possibility to alternate between the two modes.

As far as we know, the bet-observe task is the only paradigm that allows individuals to alternate between pure exploration and pure exploitation. In the bet-observe task, an individual is presented with two options. On each of M trials, one of the two options will produce a reward indicated by a green light. On each trial, the participant selects an option and makes one of two choices. He can *observe* an option, see which one produces the reward, but not receive the reward. Or he can *bet* on an option. If the player bets on an option, he will gain its underlying reward but will not see whether the reward is present. Because the participant only sees the option outcome if he observes, he can only learn about the options' underlying distributions on observation trials, but can only reap rewards on betting trials.

Navarro and Newell (2014) derived optimal decision strategies for two versions of the game: stationary and non-stationary. In the stationary version of the game, the reward probability distributions are fixed. Specifically, the probability that the left option has a reward l_p is fixed and does not change over time. In the stationary task, an optimal learner will begin the task by observing outcomes until he reaches a pre-defined information threshold. Once he reaches this threshold, he will switch to a betting strategy and will always bet on the perceived better option. In the non stationary version of the game, the reward probability distributions can change at any time. For example, with some probability α the probability l_p could change to a value drawn from a uniform distribution. In this version of the game, the optimal decision strategy alternates between observing and betting throughout the game. In other words, the actor will begin by observing for a few trials until a certain information threshold is reached, then he will switch to betting for a few trials. He will then switch to observing in order to see if l_p has changed.

However, because betting in the bet-observe task does not provide information, decision makers cannot learn anything on betting trials. This is not an inherent flaw in the paradigm - indeed, obscuring information from betting trials elegantly separates epistemic from pragmatic actions. However, because many, if not most, real-world decision tasks provide information on both exploration and exploitation trials, the bet-observe task is a poor model of most real-world decisions. From food choice to mate choice, exploitation decisions

(i.e.; consuming food or selecting a mate) will always provide information to the decision maker that it can use to update its impressions and guide future search.

In order to study how people alternate between explicit exploration and exploitation, we introduce the Peeks and Keeps task.

Peeks and Keeps

Peeks and Keeps is an extension of an N-armed bandit task that explicitly separates exploration and exploitation decisions. In the task, participants repeatedly select one of N options with *a priori* unknown underlying probability distributions over the course of M trials. On each trial, the participant selects an option and elects to either *observe* the next outcome without financial feedback, or *bet* on the outcome and receive the financial feedback. At the end of M trials, the participant is paid the sum of all sample outcomes revealed on bet trials. If he always observes and never bets, he receives no bonus. If he bets on every trial, he receives the sum total of all samples.

"Optimal" Search in Peeks and Keeps

How many peeks *should* people take when playing peeks and keeps? The answer to this question depends on two critical criteria: the specific search strategy a person uses, and the statistical environment they are in. With regards to search strategies, we assume that people use a simple "explore equally then exploit" strategy. This strategy assumes that people begin by exploring the environment equally across options using a pre-defined number of peeks. Once the person has used all of their peeks, they shift to an epsilon-greedy exploitation strategy.

In the following simulation, we will focus on the effect of statistical environments on optimal search strategies. Before going into the details of the simulation, we note that it is easy to derive trivial environments that would prescribe either the minimum (i.e.; 0) or maximum (i.e.; Infinite) peeks. An environment with options that only provide positive outcomes prescribes 0 peeks, while and environment with options that only provide negative outcomes prescribes infinite peeking. These trivial boundary conditions already suggest that there are a range of intermediate environments that prescribe intermediate levels of observation.

To reduce the strategy and environmental space in this simulation, we will make several restrictions. We assume that each environment has one option with a positive expected value, and one (or more) options with a negative expected value.

We varied three parameters in our simulation, one at the agent level, and two at the environment level.

Agent Parameters.

1. Number of Peeks: The number of peeks agents used ranged from 0 to 100 in steps of 5.

Environment Parameters.

1. Number of negative EV options: The number of negative EV options ranged from 1 to 3. Because there was always one positive-EV option, the number of total options in the environments ranged from 2 to 4.
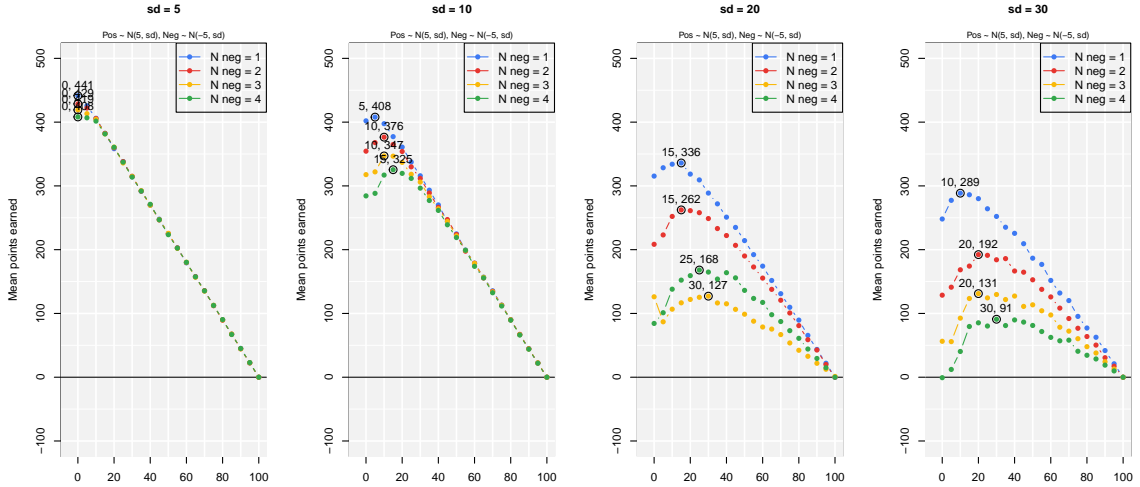
Figure 1. Aggregate results from agent-based sampling simulation. All agents began the game by peeking equally between all options until reaching a pre-determined number of peeks. Once they finished using their peeks, agents exclusively used keeps with an epsilon-greedy rule with a 0.05 probability of randomly choosing an option that was not the current best option. Each point in the plot represents the mean number of points earned by 5,000 agents playing the game with a given number of peeking trials (x-axis) in different problem environments (different panels). Each environment had one option with a positive EV (equal to +5) and one or more options (different panel rows) with negative EV (-5). All options were normally distributed; however, the standard deviation of options differed between environments (different panel columns).

2. Standard deviation of option distributions: We used three different standard deviations of option distributions: 5, 15, and 60. In each environment, the standard deviation of all options (both positive-EV and negative-EV) was the same.

    We had 5,000 agents play the game for each parameter combination. In Figure 1, we plot the environments and the median number of points earned by 5,000 agents using 0 to 100 peeks:

    The optimal number of peeks and their associated expected point earnings for each environment is presented in Table 1. Here, we see that as both the number of bad options and standard deviation of option outcomes increases, the optimal number of peeks increases. In the easiest environment, with 1 negative option and an option standard deviation of 5, the optimal number of peeks is 0 leading to an expected earning of 441.4381054 points. This suggests that this environment is so easy to learn that peeking is unnecessary, and even detrimental. In contrast, in the most difficult environment, with 3 negative outcomes and an option standard deviation of 30, the optimal number of peeks is 20 leading to an expected earning of 131.0950805 points.

n.bad	sd	npeeks	max.points	peek.MAID	peek.PB	end.MAID
1	5	0	441.44			1.18
2	5	0	428.57			1.83
3	5	0	419.12			2.26
4	5	0	408.16			2.53
1	10	5	407.85	5.12	0.86	2.04
2	10	10	376.44	4.47	0.82	2.59
3	10	10	346.99	5.13	0.72	3.11
4	10	15	325.38	4.68	0.73	3.21
1	20	15	335.95	6.00	0.83	3.44
2	20	15	262.34	7.30	0.66	4.46
3	20	30	127.10	3.83	2.84	86.71
4	20	25	168.07	7.35	0.54	5.08
1	30	10	288.58	11.00	0.69	5.62
2	30	20	192.24	9.50	0.58	6.25
3	30	20	131.10	10.93	0.47	7.14
4	30	30	90.90	9.93	0.44	6.93

Table 1

Optimal number of peeks (and resulting expected number of points) in simulation 1

Method

Participants

Participants (N = 361) were recruited from the Amazon Mechanical Turk[2]. For their participation, workers received a guaranteed reward of 50 cents with the possibility of a bonus up to $1.00. 156 (43%) were female and the mean age was (34.43) (IQR: [28, 39])

Procedure

We created four different environments by crossing two factors: Number of options (2 v. 4), and Option standard deviation (5 v. 20)[3]. The four environments are displayed in Figure 2. The order of options was randomized for each participant.

Each participant was randomly assigned to one of the 12 conditions (Response Mode (K v. PAK v. PTK) x Environment (1 v 2 v 3 v 4). In all response mode conditions, participants were told that the goal of the game was to earn as many points as possible over the course of 100 trials. To reinforce the idea that peeking introduces an opportunity cost, Those in the Peeks condition were specifically told that using a Peek action would 'use a trial.'

---

[2]We restricted our study to workers who had completed at least 100 HITs with at least a 95% HIT acceptance rate.

[3]In order to ensure that the sample distributions closely matched the desired means and standard deviations, we repeatedly generated candidate sample distributions until we found ones whose sample means were within 0.10 of the desired mean and whose standard deviations were within 1.0 of the desired value. Additionally, we truncated the distributions so the minimum and maximum values did not exceed -99 and +99 respectively.
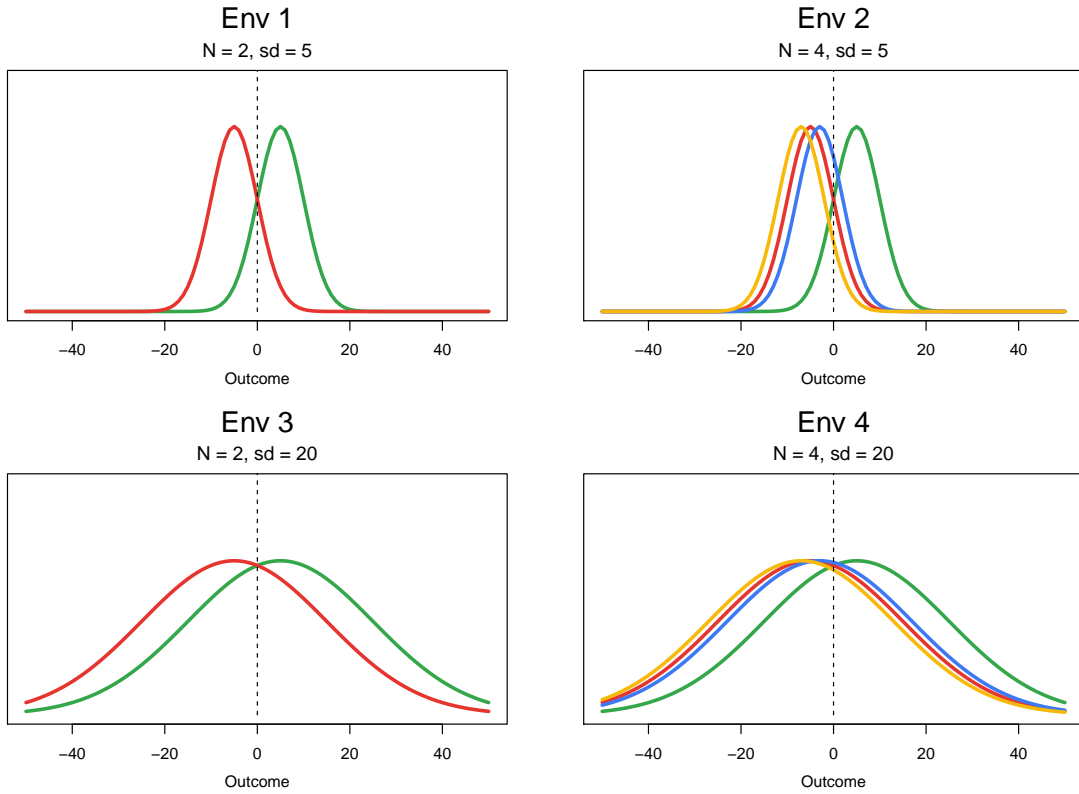
Figure 2. 4 decision environments used in Study 1. Each environment has one good option (in blue) with a positive expected value of +5. All other (bad) options have a negative expected value of -5.

After completing the first game, participants were spontaneously given a new game to play. They were told that the options in the second game were identical to those in the first game, but that the location of the options would be randomly shuffled. Participants completed three personality questionnaires (the XX, YYY, and the ZZZ) and an additional post-study questionnaire that elicited their overall impressions of the game.

Results

Final Points

Distributions of final points values across all games are presented in Figure 3.

To see which game and environmental factors affected point totals, we calculated a Bayesian ANOVA using the BayesFactor package in R.

We entered both decision environment and game type as fixed factors[4]. We found credible effects for the option environment (BF > 1000, E1–E2 = 226.07, 95% HDI [177.32, 273.68], E1–E3 = 110.08, 95% HDI [66.61, 154.71], E2–E3 = 110.08, 95% HDI [66.61, 154.71]). Critically, we found a credible effect of game.type (BF = 3: players playing PTK

---

[4]Because performance was so poor across all game types in environment 4, we elected to withdraw these data from the ANOVA. When we include these data, the effect of game type does not appear (BF = 0.14)
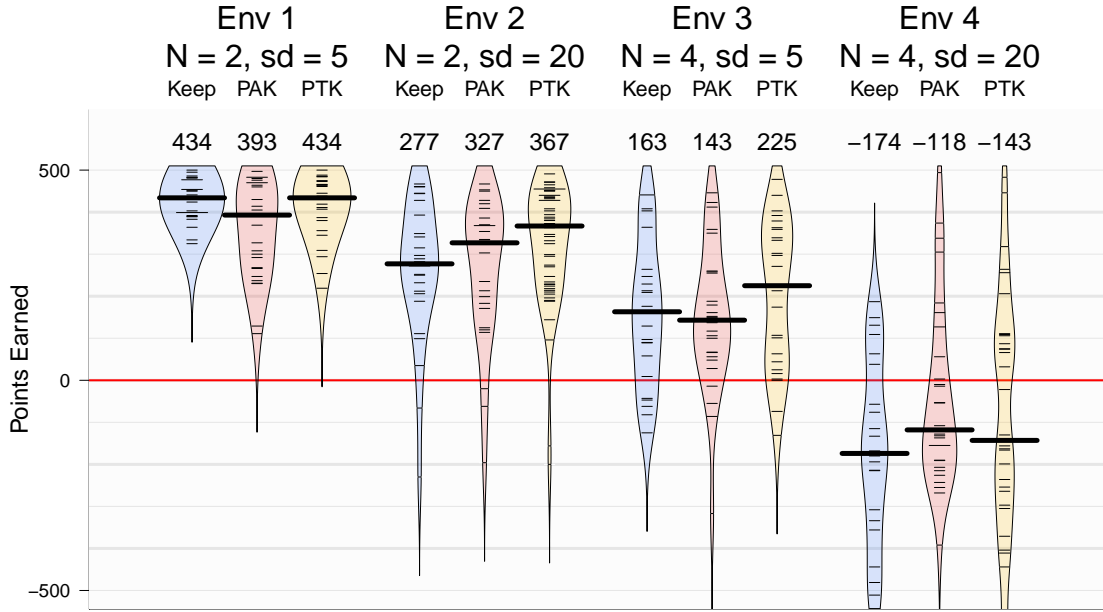
Figure 3. Distributions of final points earned by participants in each decision environment and experimental condition. Median point values are indicated by the dark black lines and are printed above each distribution.

earned more points than those playing K (mean difference = 31.36, 95% HDI [-12.25, 74.3]) and those playing PAK (mean difference = 66.35, 95% HDI [21.31, 110.97]). However, there was no credible difference between PAK and K conditions (mean difference = -34.99, 95% HDI [-79.69, 8.97]). While the raw data appeared to reflect an interaction between environment and game type on points earned, the Bayes Factor for the interaction was 0.08 indicating slight support for the null hypothesis.

Peeking Rates

Next, we focused on the two peeking conditions. We sought to answer three questions: First, how do the overall rates of peeking differ between PAK and PTK and how does peeking change over time? Second, how do environmental factors affect peeking rates? Third, in the PAK condition, how does peeking change over time? Finally, What is the relationship between option switching and peeking?

Participant-level. To see how the overall rates of peeking differed between peeking conditions, we calculated the average rate of peeking across participants at each game trial. Results are in Figure 4 and average peeking rates across trials are presented in Table 2.

- We found a credible effect of game type on number of peeks (BF > 1000: players playing PTK peeked fewer times than those playing PAK (mean difference = 12.52, 95% HDI [8.65, 16.51]).

- In PAK, peeking rates are highest in the first 10 trials. After the 10th trial, peeking rates do not change substantially over time.

Game Type	Environment	N Options	Option SD	Mean Peek Rate
pak	1	2	5	0.17 [0.16, 0.18]
pak	2	2	20	0.2 [0.18, 0.21]
pak	3	4	5	0.18 [0.17, 0.19]
pak	4	4	20	0.22 [0.21, 0.23]
ptk	1	2	5	0.07 [0.06, 0.07]
ptk	2	2	20	0.08 [0.07, 0.08]
ptk	3	4	5	0.06 [0.06, 0.07]
ptk	4	4	20	0.14 [0.13, 0.15]

Table 2

Sample mean and 95% Highest Density Intervals of peeking rates across environmental conditions

- In PTK, peeking rates start much higher than in PAK. In the first trial block of PTK, the mean peeking rate was 0.54 (95% HDI = [0.52, 0.56]). In contrast in PAK the mean peeking rate was credibly lower at 0.33 (95% HDI = [0.31, 0.35]). However, in contrast to PAK, peeking rates decreased dramatically in PTK over time.

Trial-level

In order to determine which factors were credibly correlated with the probability of peeking for participants in the Peeks and Keeps condition, we conducted a Bayesian Logistic regression analysis at the trial level. The dependent variable was a binary coded variable with a 1 indicating that the player had peeked that trial, and a 0 indicating that
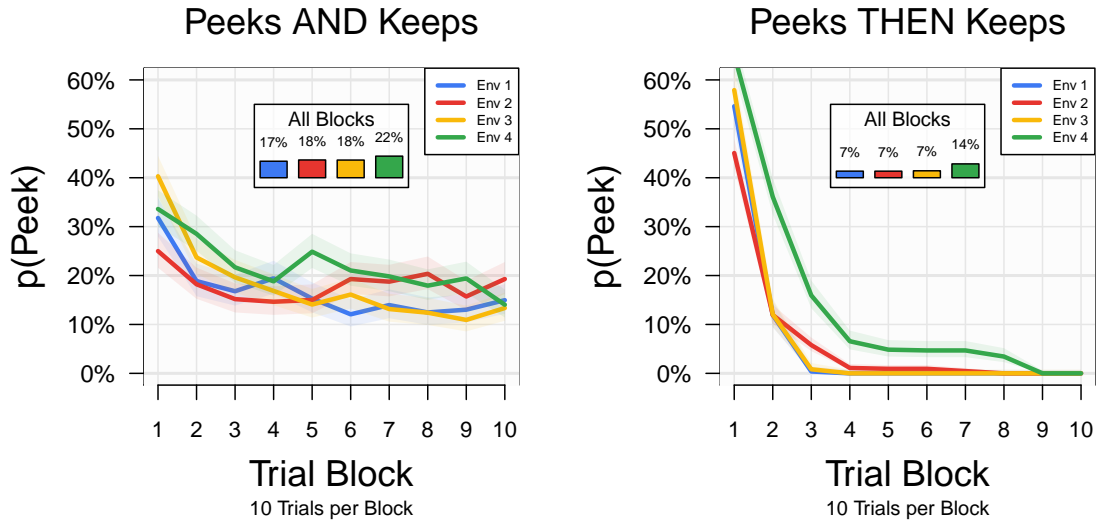

Figure 4. Mean peeking rates by trial block (groups of 10 trials) separated by environmental condition. The plot on the left is for PAK (Peeks and Keeps). The plot on the right is for PTK (Peeks then Keeps).
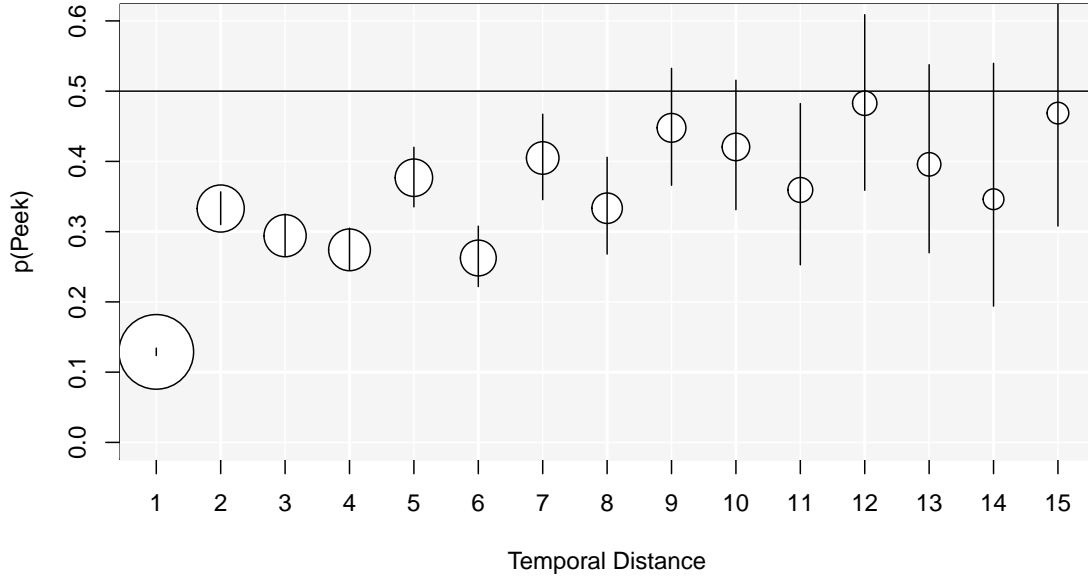
Figure 5. Mean peeking rates by temporal distance - the number of trials since the selected option was last observed by the participant - across all participants in the Peeks and Keeps conditions. The size of the symbols are an increasing function of the relative frequency of observations in the dataset. The segments indicate 95% HDIs.

the player had kept. We entered four fixed independent variables in the regression. The first two variables were time based: trial, the current trial number, and temporal distance - the number of trials since the player had last sampled from the option. The second two independent variables were a function of prior outcomes from the sampled option: prior mean positive - a binary variable with a 1 indicating that the option's prior sample mean was positive and 0 otherwise, and prior outcome positive - a binary variable with a 1 indicating that the option's prior outcome was positive.

- We begin with the time based variables. There were credible partial effect of both trial and temporal distance on peek rates. Players were credibly less likely to peek on later trials than earlier trials (-0.01, 95% HDI [-0.01, 0]), and were credibly more likely to peek as temporal distance increased (0.05, 95% HDI [0.04, 0.06]) (See Figure 5. In other words, the longer it had been since a player had sampled an option, the more likely it is that he/she peeked. For example, when temporal distance was 2, the mean peek rate was 33%, compared to a mean peek rate of 42% when the temporal distance was 10.

- Now we move to the outcome based variables. Participants were credibly more likely to peek at an option when the last outcome observed from that option was negative (29%, 95% HDI: [28, 30%]) than when it was positive (12%, 95% HDI: [12, 13%], 95% HDI of difference: [15%, 18%]).

Strategy Fitting

See pak$_s$trategy fitting.R

Study 2 (Proposed)

Design

Changes from first study

- Increase number of participants per cell (100 each = 600 participants).

- Participants are given complete information about environments prior to game. Then we can argue for 'optimal' behavior.

- Players play 3 games. Points are summed over all games and converted to a bonus. This should motivate people to maximize earnings – and minimize losses – in each game.

- Get rid of really easy (n = 2, sd = 5) and really difficult (n = 4, sd = 20) conditions.

  Between-Subjects (9 conditions)
  Game (3)

- Game Type (Keeps, Peeks and Keeps, Peeks Then Keeps)

  Environments (3)

- N.Options (2, 3, 4)

- Option SD (10)

  Within-Subjects (3)
  N-Games (3)

Discussion

Conclusion

Appendix

Additional Figures

References

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

Neth, H., & Müller, T. (2008). Thinking by doing and doing by thinking: a taxonomy of actions. In C. H. . T. S. L. Carlson (Ed.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 993–998).

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. , *71*(5), 680.

Yechiam, E., Busemeyer, J. R., & Stout, J. C. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological . . . .*