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

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.

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.

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:

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

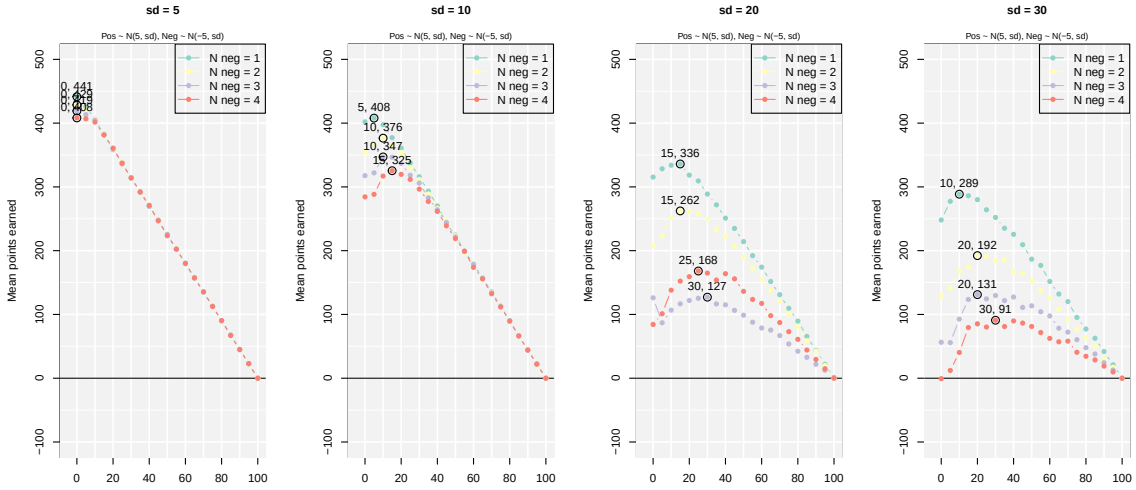

Figure 1. Aggregate results from agent-based sampling simulation. All agents began the game by peeking equally between all options until reaching a pre-determined number of peeks. Once they finished using their peeks, agents exclusively used keeps with an epsilon-greedy rule with a 0.05 probability of 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).

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 unecessary, 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.

Required learning in each environment

How much learning is necessary in each environment? To answer this, we calculated how well agents using the optimal number of peeks learned their environments. We defined learning with two measures: mean absolute impression deviation (MAID), the mean absolute difference between agent's impressions of options and the option's true EV, and prefer best (PB), the probability that, at the end of its peeking trials, the agent preferred the best option.

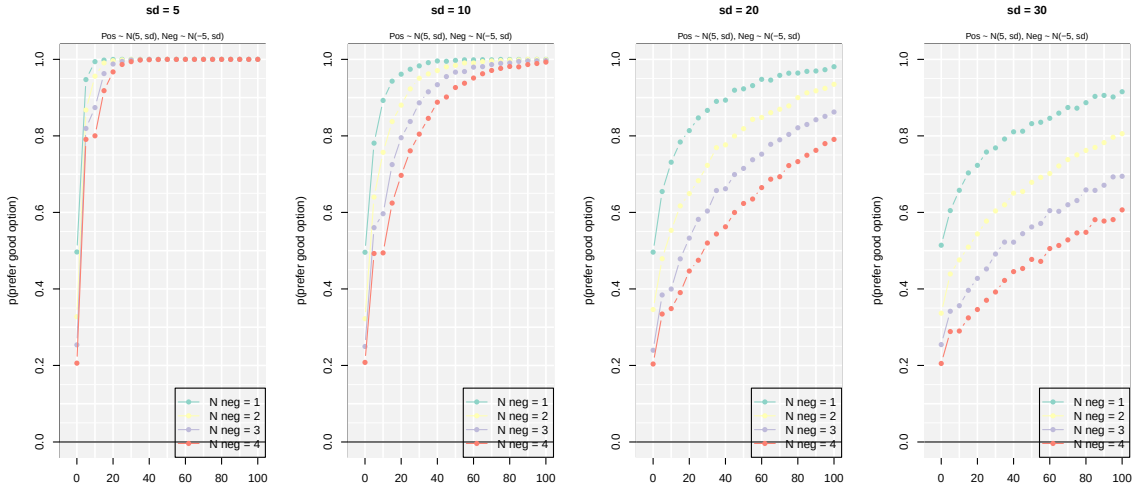

Figure 2. Aggregate results from agent-based sampling simulation. learning results.

```
## pdf
##   2
```

```
## NULL
```

Method

Participants

Participants (N = 250) were recruited from the Amazon Mechanical Turk[1]. For their paritcipation, workers received a guaranteed reward of 50 cents with the possibility of a bonus up to $1.00. 122 (49%) were female and the mean age was (35.66) (IQR: [27, 42])

Procedure

We created 6 sequences of 200 integers following a rounded Normal distribution with means of -5, 0 or +5 (corresponding to the positive, neutral, and negative options) and standard deviation of 10 or 30 (corresponding to the easy and difficult distributions)[2]. In the *dynamic* condition, the location of the positive and negative options changed places on trial 101, while in the *stable* condition there was no change. To prevent any option order counfounds, we employed all 6 possible orderings of the options on the screen as a between-subjects factor. However, for all of our analyses we have ignored this factor.

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

[2]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.

Each participant was randomly assigned to one of the 48 conditions (Response Mode (Peeks vs. Keeps) x Stimuli Difficulty (Easy vs. Hard) x Environment Stability (Stable vs. Dynamic) x Option Order). In both response mode conditions, participants were told that the goal of the game was to earn as many points as possible over the course of 200 trials using their 200 Keeps (for the Keeps condition) or their 200 Peeks and Keeps (for the Peeks condition). 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.' Participants were not explicitly told that the options would be either stable or dynamic. Instead, all participants were told that at any given point in the game, one of the options would be the best one.

After completing all 200 trials, 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

Point Totals

Summary statistics of the cumulative point totals earned by participants in each of the experimental conditions are shown in Table 3. Additionally, group mean and individual level cumulative point values across trials are shown in Figure /reffig:pointsbytrial. To see which experimental variables affected point totals, we conducted a Bayesian regression analysis with each participant's point total as the dependent variable and the three experimental conditions as independent variables. Results are shown in Table 2. We found a credible negative effect of the difficult stimuli condition for both trials 1-100 and trials 101-200, suggesting that participants did worse in the difficult environment than the easy environment. For trials 101-200 we found that participants in the dynamic condition performed credibly worse than those in the stable condition.

Condition	Trials 1 - 100	Trials 101 - 200	All Trials
Mode = Peek	-33 [-70, 17]	-34 [-81, 14]	-62 [-133, 9]
Difficulty = Hard	-109 [-153, -74]	-58 [-111, -11]	-171 [-244, -111]
Stability = Stable	30 [-4, 77]	132 [81, 179]	171 [112, 247]

Table 2

Posterior means and 95% highest density intervals of the effects of experimental conditions on point totals.

Observation Rates

Next, we restricted our analyses to the peek condition. For each participant, we calculated the percentage of trials the person peeked. In Figure 4 we show the overall distribution of participant level peek rates (across experimental conditions) and the mean peek rate across participants at blocks of 20 trials. The median participant peeked on 32% of trials: however there was a clear bimodal distribution of peeking rates wherein most participants either peeked either less than 10% of trials or between 40% and 60% of trials. Moreover, of those participants who peeked on less than 10% of trials.

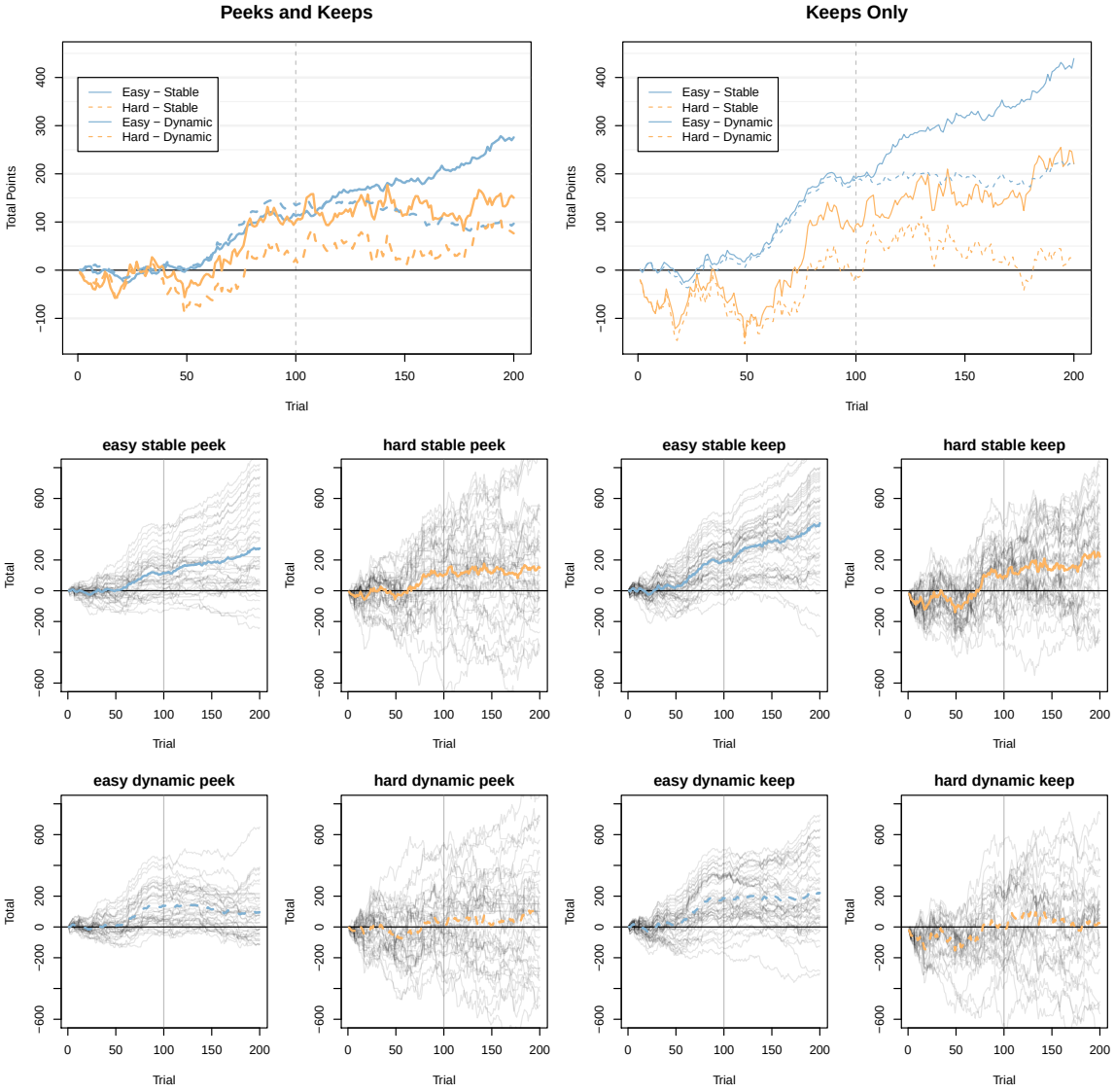

Figure 3. Cumulative point tradjectories across trials for all conditions and participants. The two panels in the top row show the average cumulative point values across participants for each experimental condition. The bottom 8 plots in the bottom two rows show the mean tradjectory of each condition separately, with individual lines plotted for each individual participant.

Mode	Difficulty	Stability	Trials 1-100	Trials 101-200	All Trials
keep	easy	dynamic	181, [120, 243]	33, [-25, 96]	211, [119, 320]
keep	easy	stable	203, [159, 256]	239, [168, 306]	445, [334, 553]
keep	hard	dynamic	26, [-84, 124]	17, [-70, 104]	45, [-82, 179]
keep	hard	stable	71, [-1, 149]	142, [56, 249]	206, [68, 346]
peek	easy	dynamic	158, [104, 206]	-21, [-72, 21]	118, [49, 182]
peek	easy	stable	112, [63, 155]	172, [99, 250]	285, [178, 380]
peek	hard	dynamic	22, [-59, 93]	48, [-62, 148]	66, [-58, 188]
peek	hard	stable	154, [52, 247]	61, [-21, 149]	194, [51, 326]

Table 3

Sample means and 95% highest density intervals of cumulative point values earned by participants within a specified trial range.

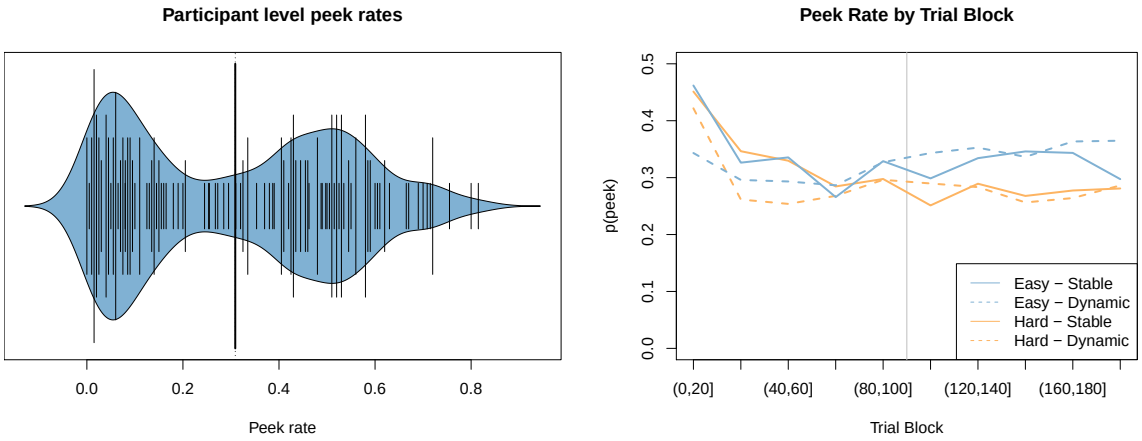

Figure 4. Distributions of peeking rates by participant (left panel) or trial (right panel).

To see which variables affected peeking rates, we conducted a Bayesian binary logistic regression with peeks at the trial level as the dependent variable, experimental conditions and trial as between-subject independent variables. We found a credible negative effect of trial number suggesting that people are less likely to peek on later trials -0.002, [-0.003, -0.002]. There were no other credible effects[3].

How did option switching relate to mode switching? If the two variables are unrelated, then selection switching may be completely unrelated to exploration. However, if the two are perfectly related, then the two measures might be psychologically equivalent. To test this, we correlated the each participant's option switching rate with their mode switching rate. The correlation was positive and credibly different from 0

The effect of observation on rewards

To see how peeks affected point totals, we regressed each participant's total points earned on the interaction between the number of peeks they took and the stability and

[3]Men were credibly less likely to peek than women: -18.49, [-35.5, -0.19]

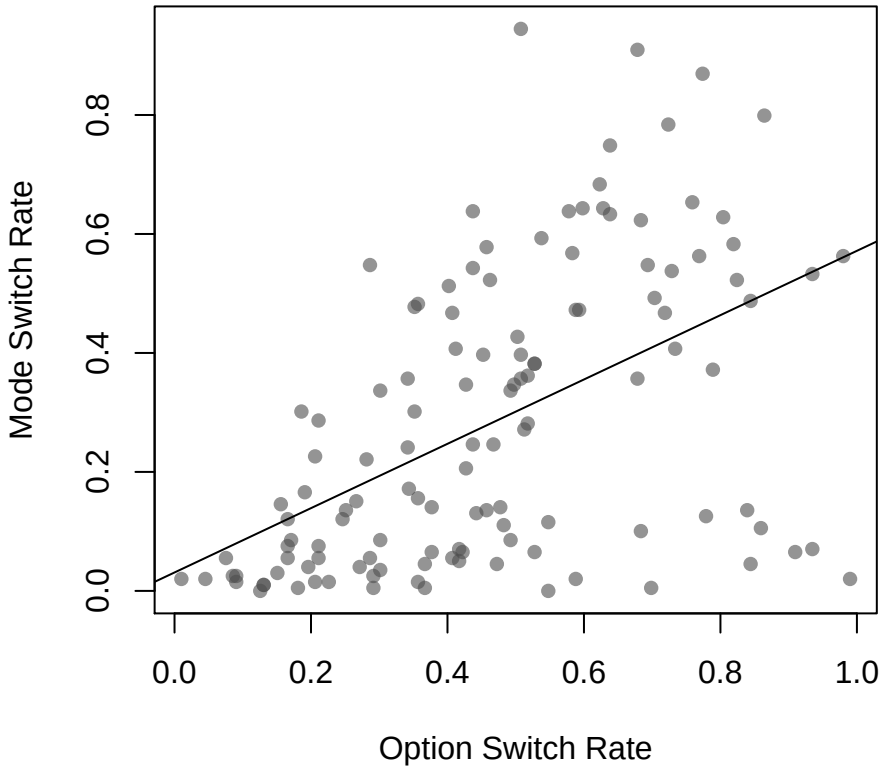

Figure 5. Distributions of peeking rates by participant (left panel) or trial (right panel).

difficulty experimental conditions. We found a credible negative interaction between the number of peeks a person took and the stability condition (-3.44, [-6.56, -0.57]), suggesting that in the stable environment, as peeks increased a person's total points decreased. The effect in the dynamic environment was negative but not credibly different from 0 (-0.88, [-3.06, 1.43]). These relationships are shown in Figure 6.

Option and Mode Switching

Did the three personality measures of impulsivity, regret, and maximizing affect search behavior? We regressed each participant's mean option switching rate and mean mode switching rate (for participants in the peeks condition) on the three personality measures. For option switching, we found credible positive effects for impulsivity (0, [0, 0.01]) and regret (0.01, [0, 0.01]), suggesting that the more impulsive and regretful a person was, the more likely they were to change options during search. For mode switching, we found a credible positive effect of maximizing (0.01, [0, 0.02]), suggesting that the more maximizing

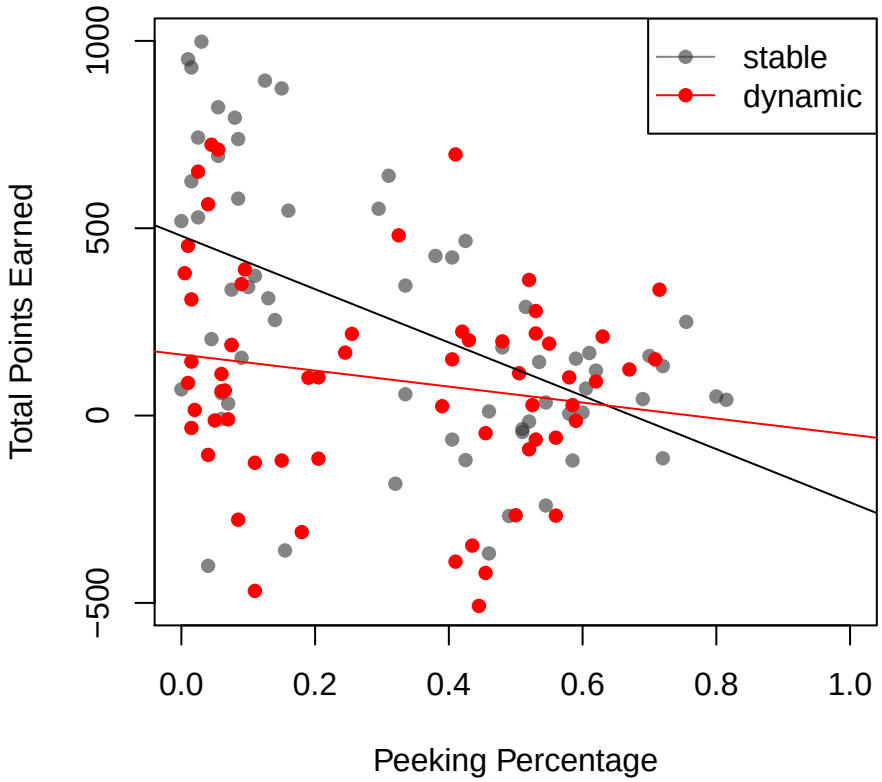

Figure 6. Relationship between number of peeks and point totals separately for the dynamic and static experimental conditions. Note To Authors: These data definitely violate homoskedasticity assumptions of linear regression.

a person was the more likely they were to change between peeking and keeping states[4].

Discussion

Conclusion

[4]We also conducted a similar regression analysis with each person's total percentage of peeks as the dependent variable. None of the effects were credibly different from 0.

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 ...*.