Does present bias influence exploratory choice?

Alexander S. Rich Todd M. Gureckis

Does a present bias influence exploratory choice?

Decision makers that act in uncertain environments frequently face dilemmas between exploiting options known to be rewarding and exploring options that are uncertain. A kid buying ice cream, for example, must select between getting a cone of her favorite flavor and trying something new that might become a new favorite but could also be disappointing. Researchers in reinforcement learning have created a large body of knowledge about how people and animals handle the explore–exploit dilemma (Mehlhorn et al., 2015) and how the problem should be approached computationally (Sutton & Barto, 1998).

A key aspect of exploratory decision making is that it is spread over time. With a single decision, exploration makes little sense. If, heaven forbid, you only have one remaining chance to buy ice cream in your life, you should buy your favorite flavor, as that is the flavor you expect to enjoy the most. Exploring new flavors has the possibility of introducing you to a new favorite, but it is only when there will be many more chances to get that new flavor in the future that the risk of a disappointment starts to look worthwhile.

While research on exploratory choice acknowledges that the value of exploration depends on its payoffs in the future (Rich & Gureckis, 2017; Wilson, Geana, White, Ludvig, & Cohen, 2014), it has not addressed how this might interact with the manner in which people value future rewards. When considering rewards spread over time, people tend to be present biased, overweighting immediate rewards in comparison to delayed rewards (Frederick, Loewenstein, & O'Donoghue, 2002; Myerson & Green, 1995). In scenarios with explore–exploit tradeoffs, this could lead to over-exploitation and under-exploration. While lab experiments tend to be conducted in short sessions with non-consumable rewards, preventing present bias from being a major factor, this preference for immediate reward could be a major factor leading to under-exploration in more temporally extended, real world settings.

In this paper, we discuss the potential connection between exploratory choice and intertemporal choice. We report on a set of experiments using directly consumable rewards in an exploratory choice task to test for effects of present bias on exploration. While we did not find an effect of present bias on exploratory choice, a follow-up experiment revealed that our consumable rewards did not in fact produce a reliable present bias, despite evidence that they did so in earier studies (D. Navarick, 1998; Solnick, Kannenberg, Eckerman, & Waller, 1980). Nonetheless, we hope that this work can serve as an interesting first step towards unifying our understanding of exploratory and intertemporal decision making.

Exploration inside and outside the lab

Many researchers have examined patterns of exploration both in naturalistic settings and in the lab. Interestingly, differing findings have emerged as to the nature and severity of biases in exploratory choice.

Exploration outside the lab

Exploration has been studied outside the lab in a wide range of contexts. While these domains vary greatly in their superficial characteristics, a bias towards under-exploration has often been observed.

Learned helplessness, a phenomenon applicable to many behaviors and domains, has been described as an example of insufficient exploration. In learned helplessness, an organism experiences the absence of control over the environment, learns that the environment is uncontrollable, and thus ceases to take actions that might allow it to discover that it can in fact exert control. Learned helplessness has been proposed to underly some forms of depression (Lyn Y Abramson, Metalsky, & Alloy, 1989; L Y Abramson, Seligman, & Teasdale, 1978) as well as problems ranging from difficulties in school (Diener & Dweck, 1978) to poverty (Evans, Gonnella, Marcynyszyn, Gentile, & Salpekar, 2005). While the cognitive appraisal of experienced events affects the development of learned helplessness (L Y Abramson et al., 1978), patterns of exploration clearly play a role as well (Huys & Dayan, 2009, Teodorescu & Erev (2014a)). In the case of depression, interventions aimed at increasing the exploration of activities that might be rewarding have been found to be as effective as those with a more cognitive orientation (Jacobson et al., 1996).

Under-exploration also seems to occur in the development of complex skills, such as flying a plane or playing a sport (D. Gopher, Weil, & Siegel, 1989). In these settings, an "emphasis change" training method that encourages people to continually explore the performance space leads to greater performance gains than unguided practice or more complex training methods. Without this intervention, people often enter a "local maximum" in which exploration decreases and performance plateaus (Yechiam, Erev, & Gopher, 2001).

In many other areas under-exploration is less clearly established, but is suspected to play a role in maladaptive behavior. Insufficient exploratory interaction with outgroups may be one cause of stereotypes and prejudice (Denrell, 2005), and interventions that increase inter-group contact reduce stereotypes (Shook & Fazio, 2008). The crowding out of exploration by exploitation is a concern in organizational behavior as well (Levinthal & March, 1993; March, 1991), prompting research into organizational structures that may preserve exploration (Fang, Lee, & Schilling, 2010).

Exploration inside the lab

Lab studies of exploratory choice have allowed researchers to fully control the reward structure of the environment and precisely measure behavior, as well as compare behavior to optimal choice and other formal models. These studies have yielded a number of insights into the factors leading to more or less exploration, including aspiration levels (Wulff, Hills, & Hertwig, 2015), uncertainty (Speekenbrink & Konstantinidis, 2015), and the future value of information (Rich & Gureckis, 2017; Wilson et al., 2014)

Interestingly, under-exploration has not emerged as a clear pattern in lab experiments. Instead, results are mixed with people sometimes under-exploring, sometimes over-exploring, and sometimes exploring close to an optimal amount. To take two illustrative examples, (Zwick, Rapoport, Lo, & Muthukrishnan, 2003) found that in a sequential search task people under-searched when there were no information costs but over-searched when there were information costs, and (Teodorescu & Erev, 2014b) found that people explored unknown alternatives too often or not often enough depending on whether rare outcomes were positive or negative. Similar results have been obtained within and across a variety of other studies and paradigms (Hertwig, Barron, Weber, & Erev, 2004; Juni, Gureckis, & Maloney, 2016; Navarro, Newell, & Schulze, 2016; Sang, Todd, & Goldstone, 2011; Tversky & Edwards, 1966).

These experimental studies raise the question of why under-exploration appears more widespread in the field, but not in the lab. One possibility is that both forms of deviation from optimality are in fact prevalent, though perhaps in different settings, and that the seemingly general bias toward under-exploration is illusory. An alternative is that there are some important aspects of real-world decisions—or peoples' cognitive and motivational states when making those decisions—that makes differentiate them from decisions in the lab. One clear possibility is that in real world exploration, choices and outcomes are spread out over time in a manner that is rarely found in the lab, and that people's bias towards immediate rewards might therefore account for a portion of people's tendency to under-explore.

Temporal discounting

Temporal discounting refers to the underweighting of temporally distant rewards relative to close ones, and is a ubiquitous phenomenon across decision-making agents including people, animals, and organizations. Temporal discounting is rational if it occurs at an exponential rate δ, where the value of a reward r at time t is

$$V(r,t) = re^{-t\delta}$$

In exponential discounting, each additional unit of waiting time decreases the value of a reward by an equal proportion (Samuelson, 1937, Frederick et al. (2002)). This means that the relative values of an early and a late rewards are the same no matter what time point they are considered from, or equivalently that their relative values are unaffected by adding an additional waiting time to both.

An extensive literature documents that people and animals violate exponential discounting. Specifically, in the short term rewards are discounted at a steep rate with each additional unit of waiting time, while in the long term rewards are discounted at a shallow rate. This sort of non-exponential discounting leads to a present bias, in which in the short term people excessively over-weight immediate over future rewards. For example, people will often prefer a larger, later monetary reward to a smaller, sooner reward when both rewards are in the future, but will switch their preference when the time until both rewards is reduced so that the sooner reward is immediate or nearly immediate (Kirby & Herrnstein, 1995). With monetary rewards, the delay or speed-up required to observe preference reversals is usually several days. With non-monetary rewards, such as the cessation of an annoying noise (Solnick et al., 1980), watching a video when bored (D. Navarick, 1998), or drinking soda when thirsty (Brown, Chua, & Camerer, 2009), a bias towards immediate rewards has been observed on the scale of minutes or even seconds.

There is debate about how to formally describe non-exponential discounting. Many studies have found that humans and animals appear to discount future rewards at a hyperbolic rate, allowing the value of a future reward to be written

$$V(r,t) = \frac{r}{1+kt}$$

for appropriate constant k (Myerson & Green, 1995). This formulation often fits data well, but is difficult to deal with analytically. An alternative is to posit that discounting after the present proceeds exponentially, but that there is a one-time drop in value when the reward goes from being immediate to being in the future (D. Laibson, 1997). In this model, known as the beta–delta or quasi-hyperbolic model, the value of a future reward is

$$V(r,t) = \begin{cases} r & \text{if } t = 0 \\ \beta r e^{-t\delta} & \text{if } t > 0 \end{cases}$$

where δ is the rate of exponential discounting and β is the degree of present bias. This model suffers from ambiguity in when exactly the "present" ends and the future begins. (E.g., should the value of reward received in 30 seconds be discounted by β, or should it be considered immediate?) However, it captures in a simple and tractible way many of the qualities of human intertemporal choice, and for this reason we will adopt it for our additional analyses below.

Exploration and temporal discounting

The rewards from exploratory choice are inherently distributed over time. In expectation, an exploitative action yields the greatest reward in the present, because it is the action *currently believed* to yield the highest

reward. An exploratory action is expected to yield less immediate reward, but it can compensate for this by providing useful information. This information can allow the decision-maker to make better choices in the future, leading to higher rewards later on.

Thus, temporal discounting plays a central role in determining the balance between exploration and exploitation. Rational, exponential discounting ensures that a decision-making agent explores neither too little nor too much given its degree of interest in the future. Some degree of discounting is generally good, because at some point the distant gains from continued exploration are not worth their immediate costs (Le Mens & Denrell, 2011). But as past theoretical work has highlighted, discounting that is too steep or that exhibits a present bias can lead to chronic over-exploitation and under-exploration (March, 1991, Levinthal & March (1993)).

To understand how patterns of discounting affect exploration, consider a simple scenario in which an agent must make a sequence of choices between two actions. Action A is to choose a sure-bet option that always provides a payoff of 2. Action B is to choose from a large set of uncertain options. Each uncertain option has a 25% chance of producing a payoff of 4, and a 75% chance of producing a payoff of 0. Once a high-payoff uncertain option is found, it can be selected on every subsequent choice.

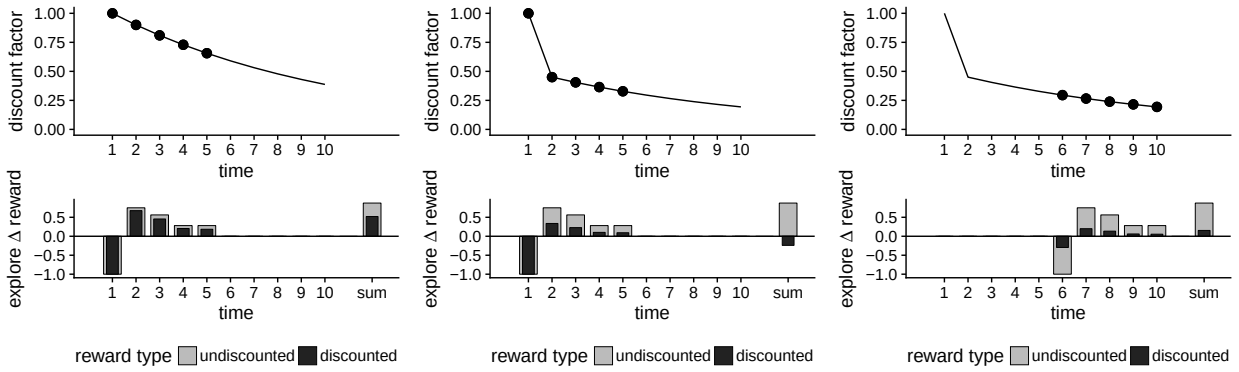

Figure 1: Effects of exploration over time for different discount curves in a simple exploratory choice task (see text for more details). The top row of panels show the degree of discounting at each time step. The bottom row of panels show the expected change of undiscounted (gray) and discounted (black) reward at each time step from exploring at the *first* action. The left panels shows exponential discounting, the center panels show quasi-hyperbolic discounting, and the right panels show quasi-hyperbolic discounting with a front-end delay. Exploration appears worthwhile to an agent with exponential discounting or quasi-hyperbolic discounting with a delay, but not to an agent with quasi-hyperbolic discounting and no delay.

This scenario presents an explore exploit dilemma because as long as a high-payoff option has not been found, the best immediate action is A, with an expected payoff of 2, rather than action B, with an expected payoff of $.25 \cdot 4 = 1$. Long term payoffs, in contrast, are increased by exploring the options available through action B, because the agent may find a high-payoff option that can be exploited on all future choices.

Whether the agent decides to forgo the immediate gains of exploiting A in order to explore B will depend on how much it values the future. Figure 1 shows the effects of various patterns of discounting on the expected rewards over a sequence of five choices. The left column shows the case of exponential discounting. The top graph shows the exponential discounting curve, with dots indicating the time and weight of each of the five choices. The bottom graph shows the change in expected reward at each choice that is caused by selecting action B rather than A at the *first* choice. (This analysis assumes that all subsequent choices are made optimally in terms of undiscounted rewards.) For mild exponential discounting, we see that at time 1, choosing B over A causes a steep decrease in expected reward, because it trades an expected payoff of 2 for an expected payoff of 1. At times 2–5, however, the expected payoff goes up; choosing B at time 1 can only increase payoffs at later times, by revealing an high-payoff option. At the far right of the graph, we see that summed discounted reward, in black, is positive, and thus that the agent will choose to explore. The undiscounted reward, in gray, is larger, but doesn't differ in sign from the reward after mild exponential discounting.

The center column shows the case of beta–delta, or psuedo-hyperbolic, discounting. As the top graph shows, rewards from later time points are weighted much less than in exponential discounting. Because of this, the expected gain in future reward for choosing B becomes smaller, while the immediate expected loss remains the same. The summed discounted reward becomes negative, and the agent adopts a completely exploitative policy of choosing A instead of initially exploring the uncertain action B.

To preview our experimental manipulation, the right column shows a case of beta–delta discounting considered from a temporal distance. Now, the first choice is at time 6, while the last is at time 5. Suppose the agent was given the opportunity to commit to a first action from time 1. As shown in the bottom graph, the sequence of rewards viewed from this distance is highly discounted, but is no longer heavily biased towards the first outcome. Instead, the expected discounted reward sequence is now a scaled version of the exponentially discounted rewards, since beta–delta discounding is identical to exponential discounting after the present. The summed expected rewards for exploration are greater than those for exploitation, so the agent will choose the exploratory action.

Capturing present bias in exploratory choice

As alluded to earlier, several approached have been used to study present bias in the lab. Many studies use monetary rewards, and offer participants various one-off choices between different quantities of money at different delays to determine their discounting curve (Myerson & Green, 1995). However, exploratory choice is inherently not "one-off." Choices can only be considered exploratory or exploitative if they are embedded within an ordered sequence of choices, where the knowledge gained from one choice can be used to inform the next. Thus, to study present bias during exploratory choice, an experiment must include a sequence of choices and outcomes, with enough time between them for discounting of the future to be non-negligible. With monetary rewards, this means the choices in an experiment would have to be spread out over weeks or months. This leads us to consider non-monetary, directly consumable rewards.

While people tend to discount money relatively slowly, they often discount primary rewards significantly for delays of minutes or seconds. This can be measured in a number of ways. In some cases, an explicit choice between a larger later and a smaller sooner reward is offered. Mcclure, Ericson, Laibson, Loewenstein, & Cohen (2007), for example, found that thirsty participants showed present bias when asked to choose between larger and smaller juice rewards separated by a few minutes. In other cases, the choice between a smaller-sooner and larger-later reward is offered repeatedly, but without explicit description, and participants are allowed to build a preference through experience. Using this, researchers have found present biases on the scale of seconds for playing a video game, watching a movie, or relief from an annoying noise (Millar & Navarick, 1984; D. Navarick, 1998; Solnick et al., 1980).

In the above studies, each choice is "one-off," creating rewards but not affecting future choices. Brown et al. (2009) provided evidence of present bias in a task in which immediate consumption affected consumption from future choices. They created a life-cycle savings game in which participants gained income and decided how much to spend over 30 periods spaced a minute apart. They arrived to the experiment thirsty, and were allowed to consume their spent income in the form of soda. In the immediate-reward condition, participants made choices at each period and then immediately consumed their soda reward. In the delayed-reward condition, the experimenters imposed a 10 minute delay between choices and reward consumption; thus, after a choice was made, the soda earned from that choice was consumed 10 periods later. Participants in the delayed-reward condition were able to consume more total soda on average, suggesting that the temporal delay decreased their present bias and allowed them to choose in a manner leading to greater long-term reward.

In the following two experiments, we use an intervention similar to that of Brown et al. (2009) to test for effects of present bias on exploratory choice. As indicated in Figure 1, if an exploratory choice task is paired with immediate consumption we predict present bias to lead to underexploration. However, if a temporal delay is introduced between decisions and outcomes, the present bias will be decreased, leading to greater exploration.

We used videos as a positive outcome that could produce present bias (D. Navarick, 1998), and a boring slider task (Gill & Prowse, 2012), along with, in Experiment 2, annoying noises (Solnick et al., 1980), as negative outcomes. It is worth noting that we also piloted an experiment using a video game as a positive outcome (Millar & Navarick, 1984), but found that participants did not find the video game sufficiently enjoyable. Experiment 1 represents a first attempt to test for effects of present bias on exploratory choice, and Experiment 2 is a larger, preregistered study that improves on Experiment 1 in several ways. After finding no evidence of present bias producing an effect in Experiments 1 or 2, in Experiment 3 we test directly, using a simpler design, whether our outcome stimuli in fact produce a consistent preference towards immediate rewards.

Experiment 1

Methods

Participants

Forty participants completed the experiment, which was conducted over Amazon Mechanical Turk (AMT) using the psiTurk framework (Gureckis et al., 2015). Participants were paid $5.00 for their participation, with a performance-based bonus of up to $3.00. All participants received the full $3.00 bonus. Participants were pseudo-randomly counterbalanced across two conditions.

Design and procedure

Consumption tasks

Participants were informed that their job was to perform a monotonous slider task that would be split into 30-second "work periods," but that they would be able to make choices throughout the experiment that would give them a chance to watch a YouTube video instead. The number of remaining work periods in the experiment was shown at the top of the screen, as was the number of seconds left in the current work period.

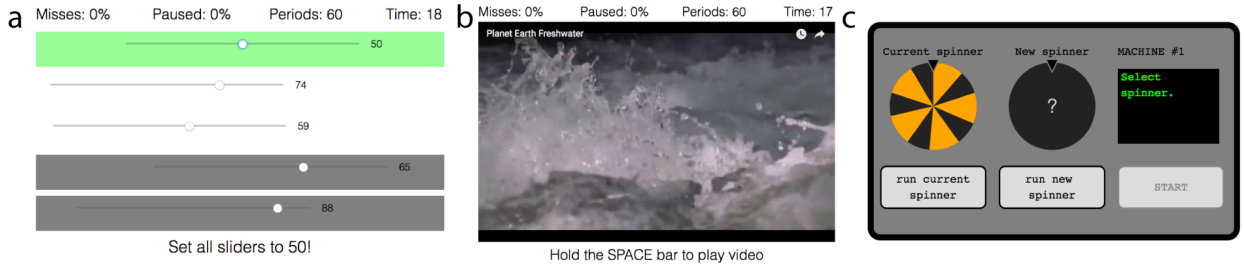

Figure 2: Examples of the Experiment 1 tasks. (a): an example of the slider task. Participants had to move all sliders to "50" in 30 seconds. (b): the video-watching task. Participants had to hold the space bar to watch their chosen video. (c): the decision-making task. Participants had to choose to run the machine with the current spinner or try a new spinner. If their chosen spinner landed on a gold wedge, they performed the video-watching task instead of the slider task.

The slider task was based of a task previously used by (Gill & Prowse, 2012). In each period of the slider task, five horizontal sliders appeared on the screen (Figure 2a). Each started at a random setting between 0 and 100, with the slider's value shown to its right, and with a random horizontal offset so that the sliders were not aligned. The participant's task was to use the mouse to move each slider to "50" before the work period ended. When a participant released the mouse at the correct setting, the slider turned green to show it had been completed. To ensure that the task took close to the allotted 30 seconds, at the beginning of the

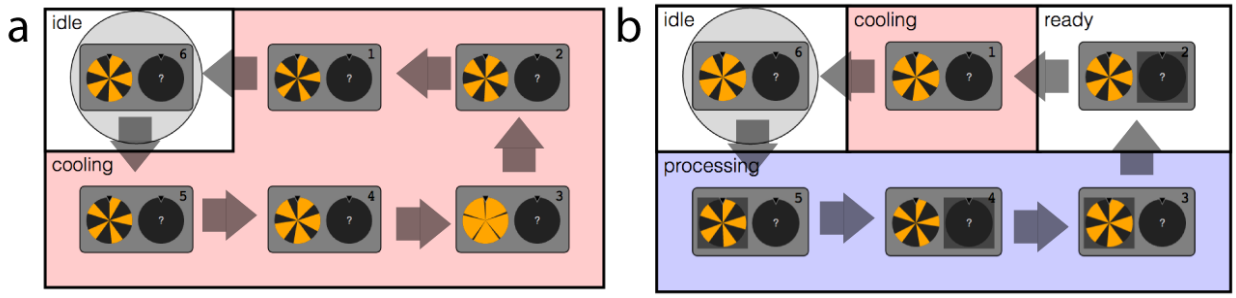

Figure 3: The machine display seen by participants. The display allowed participants track the value of each machine and the next time each machine would be ready to make a choice or produce an outcome. Gray arrows have been added to depict the counterclockwise movement of machines around the display after each work period. (a): the display seen by participants in the immediate condition of Experiment 1. (b): the display seen by participants in the delayed condition of Experiment 1.

task only the top slider was enabled, and the other four were grayed out. Additional sliders were enabled at five-second intervals, such that all five sliders were available after 20 seconds.

Before beginning the experiment, participants chose one of four videos available on YouTube: an episode of "Planet Earth", and episode of "The Great British Bakeoff", and episode of "Mythbusters", or an Ellen Degeneres comedy special. The video was embedded in the experiment with all user controls (such as skipping ahead) disabled (Figure 2b). When given access to the video, participants had to keep the browser window open and hold down the space bar for the video to play. This allowed us to ensure that participants maintained engagement with the content.

Participants completed a total of 70 work periods. For the first 10 work periods, participants simply clicked a button to begin the slider task. After these initial periods, participants gained access to six machines that could potentially complete the slider task for the participant, allowing the participant to watch their chosen video instead. However, the machines did not always function, and participants had to make a decision about how to set the machine before each use.

Decision-making task

Following the initial 10 periods, participants were shown a machine before each work period and, as shown in Figure 2c, had to select between two circular spinners with arrows at the top: the "current spinner" (exploit) and the "new spinner" (explore). The current spinner was split into five black and five gold wedges. If a participant chose the current spinner, it spun and, if the arrow landed on gold, the machine worked and the participant could watch the video. Initially, the current spinner's gold wedges were randomly set for each machine to cover between 1/3 and 2/3 of the spinner.

The new spinner initially showed a question mark. If a participant chose the new spinner, then a new spinner was created and appeared on the machine. The new spinner's gold wedge could cover anywhere from 0% to 100% of the spinner. The new spinner then spun and, if the arrow landed on gold, the machine worked.

Critically, if the new spinner had a greater gold area than the current spinner, the new spinner was "saved" and the current spinner was updated to the new spinner. This created an explore–exploit tradeoff in which choosing a new spinner carried immediate risk, but could carry long-term benefits by improving the current spinner from its initial value.

The experiment's two conditions differed in what occured after the participant spun the spinner. In the immediate condition, the machine ran immediately after the choice was made and affected the next work period, as shown in Figure 3a. It then "cooled off" for the following five periods, as choices were made with the other five machines. In the delayed condition, each machine was presented to the participant four work periods before it was scheduled to run, and the participant made a choice at that time. The machine then had

to "process" for four work periods, thus delaying the outcome by over 2 minutes (Figure 3b). The participant then returned to the machine to observe its outcome and either perform the slider task or watch the video. The machine then cooled off for a single period before being ready for another choice.

Finally, in order to induce exploration throughout the entire experiment, the six machines would occasionally "reset" after they ran. When this occurred, the current spinner would be set to a new random value between 1/3 and 2/3 gold. Participants were informed that this would occur randomly on 1/6 of trials. In fact, the procedure was designed so exactly one of the six machines would reset on each cycle through the machines, and no machine would be reset on two consecutive uses.

Training, incentives, and post-experiment questions

Before beginning the full experiment, participants completed two practice phases. First, they performed several trials of practice using the machines, with the actual work periods removed. Then, they performed two work periods practicing the slider task and one work period practicing the video task. During the machine choices, participants had access to an "info" button at the bottom of the screen that provided reminders about the dynamics of the task.

Participants were given a performance-based bonus of $3.00 for completing the consumption tasks accurately. If they missed fewer than 10% of sliders throughout the experiment and left the video paused less than 20% of the time, they were not penalized. However, if they missed more sliders or left the video paused for longer, they lost 10 cents from their bonus for each additional percentage of sliders missed or time with the video paused. The running percentage of sliders missed and video pause time was displayed at the top of the screen throughout the experiment.

Following the experiment, participants were asked to rate their enjoyment of the slider task and of the video-watching task on a 1 to 7 scale.

Results

Participants rated the video as more enjoyable on average (5.65 out of 7) than the slider task (3.13 out of 7), $t(39) = 9.26$, $p < .001$.

To analyze participants' trial-by-trial decision-making, we conducted a hierarchical Bayesian logistic regression using the Stan modeling language (Stan Development Team, 2015). This approach allowed us to estimate population-level effects of the current-spinner value and of condition, while also allowing for individual differences. The regression model included an intercept term as well as terms for the value of the current spinner, the participant's condition, and a condition by value of current spinner interaction. We included predictors for the value of the current spinner, the participant's condition, and the interaction between condition and current spinner value. Condition was coded as -1 for the immediate condition and 1 for the delayed condition; current spinner value was rescaled to have zero mean and unit variance across participants. We assumed that individuals could vary in their overall tendency to explore (i.e., intercept) as well as their responsiveness to current spinner value (slope). Participants' individual-level parameters were assumed to be drawn from a t distribution with $df = 5$, making our population level estimates robust to potential outliers. The priors on the the population-level predictor coefficients, and on the standard deviation of the t distributions from which individual-level parameters were drawn, were (truncated) normal distributions with a mean of 0 and a standard deviation of 5.

The model posterior was estimated using the Stan modeling language (Carpenter et al., 2017). We ran four chains of Hamiltonian Monte Carlo sampling, with 1000 samples per chain, the first half of which were discarded as burn-in. We confirmed convergence using the $\hat{R}$ convergence criterion (A. Gelman, Carlin, Stern, & Rubin, 2014). In the results below, we report 95% credible intervals (CIs) on model parameters of interest. An overview of the model posterior is displayed in Figure 4.

Participants were less likely to choose a new spinner when the current spinner has a high value, $CI = [-4.13, -2.52]$. However, in this experiment we found no evidence of an effect of condition, $CI = [-.58, .73]$.

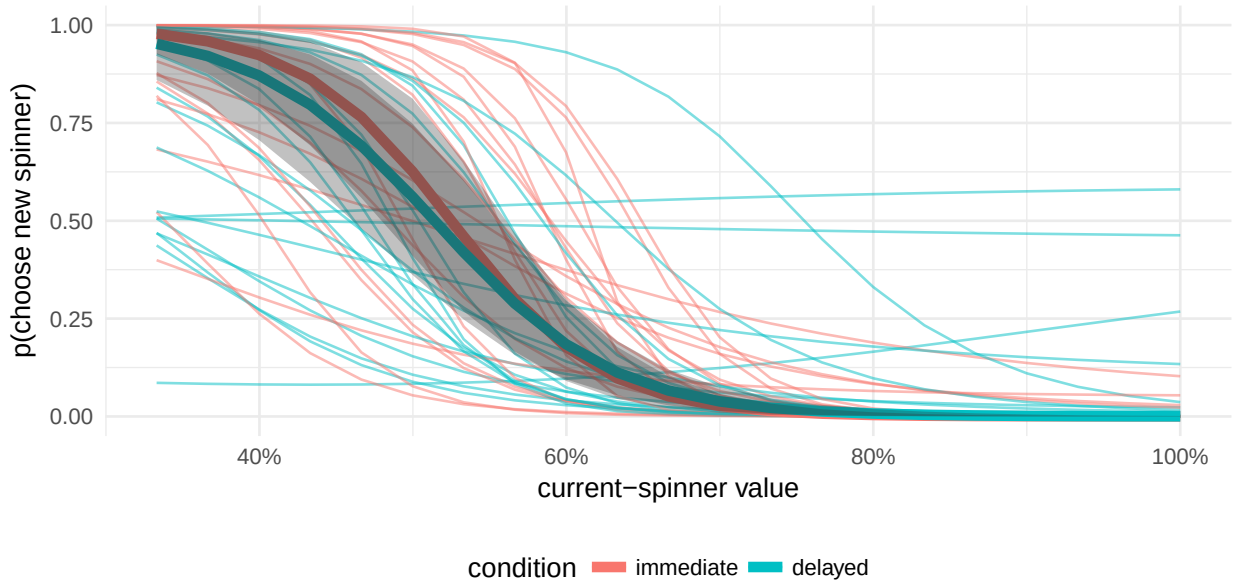

Figure 4: Model-based estimates of participants' probability of choosing a new spinner for different values of the current spinner in Experiment 1. Thick lines and shaded regions indicate the mean and 95% posterior interval for the population-level parameters, while the thin lines indicate the mean posterior parameters for each of the 40 individual participants. Participants in the delayed-outcome condition were no more likely to explore at a given current-spinner value than those in the immediate-outcome condition.

This means that participants were no more likely to explore a new spinner when there was a temporal delay imposed between their choices and the received outcomes. However, there may have been a small interaction between current spinner value and condition, such that participants in the delayed condition were less sensitive to the value of the spinner when making their choices $CI = [-.27, 1.42]$. This might indicate that the delayed condition was confusing to some participants, as a few individuals (as seen in Figure 4) changed their behavior very little across current-spinner values.

Experiment 2

In Experiment 1, we found no evidence of the delay in rewards leading to an increase in exploratory choice. However, there were several potential flaws in the experiment design which may have prevented present bias from occurring or its effects from being observed. In Experiment 2, we preregistered the design, collected a larger sample, and attempted to improve on Experiment 1 in several ways.

We conducted Experiment 2 in person, rather than using AMT. This ensured that participants had few distractions from the consumption tasks, potentially increasing their motivational effect. We also made the slider task more aversive and the video task more pleasant. To do so, we added an intermittent static noise during the slider tasks, and allowed people to switch among the four videos at will, without having to hold down the space bar to keep the video playing.

To simplify and improve the exploratory choice task, in Experiment 2 there was a single machine, rather than six. Rather than the machine "processing" for four trials in the delayed condition, outcomes were added to a "work queue" that delayed the consumption task by eight trials. This was both simpler and increased the delay length. The visual appearance of the exploratory choice task was also redesigned to make the statistics of the task more transparent.

Finally, we measured participants' impulsivity, a potentially important covariate, using the Barratt Im-

pulsiveness Scale (Patton, Stanford, & Barratt, 1995). There is evidence that this scale correlates with present-focused behavior in repeated choice tasks (Otto, Markman, & Love, 2012), though other studies have not found a relation (Brown et al., 2009).

Methods

The experiment was preregistered through the Open Science Framework. The preregistration can be found at: osf.io/3r9ke.

Participants.

One hundred people from the general community took part in the study in person at New York University. The participants had a mean age of 23.9 (SD=6.1). Fifty eight self-reported female, forty one male. Participants received $10 for taking part in the study, which lasted approximately one hour, and received a performance-based bonus of up to $5. All but one participant received a bonus of $5, with the remaining participant receiving $4.4. The experiment was approved by the Institutional Review Board at New York University. Participants who failed a post-instructions questionnaire more than twice were excluded from further analyses. Ten participants were excluded in this manner.

Design and procedure.

Barratt Impulsiveness Scale

Prior to reading the experiment instructions, participants completed the 30-item Barratt Impulsiveness scale (Patton et al., 1995) on the computer.

Consumption tasks.

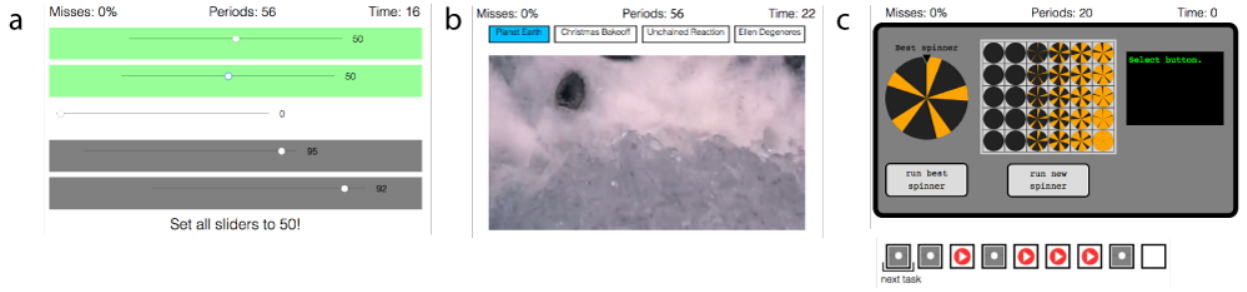

Figure 5: Examples of the Experiment 2 tasks, which resemble the Experiment 1 tasks. (a): the slider task. (b): the video task. (c): the decision-making task. After making a choice in the decision-making task, the produced outcome was added to the work queue, pictured at the bottom of (c).

Participants were informed that there were two types of tasks, a "slider task" and a "video task," that they would complete during 30-second "work periods." The number of remaining work periods in the experiment was shown at the top of the screen, as was the number of seconds left in the current work period.

The slider task was the same as the one described in Experiment 1, and is pictured in Figure 5a. To make the slider task more unpleasant, a short static noise was played through the computer speakers intermittently at a moderate volume during the task.

As in Experiment 1, the video tasks consisted of simply watching one of four videos: an episode of "Planet Earth", and episode of "The Great British Bakeoff", and episode of "Unchained Reactions", or an Ellen Degeneres comedy special. Participants watched the video through a player on the computer screen. Unlike

in Experiment 1, they did not have to hold down a button to play the video. They were free to fast forward or rewind the video at will, and could also switch among the videos at any time by clicking one of four tabs above the player (see Figure 5b).

To incentivize participants to attend to and perform the slider task, they were penalized if they missed more than 10% of the sliders. For each percentage over 10% of sliders that were not set to 50 over the course of the experiment, $.20 was deducted from a bonus that started at $5.00.

Choice task.

Participants completed a total of 56 work periods. The first eight were automatically spent performing the slider task. For the remaining 48, participants made a choice prior to each work period that determined whether the work period would be devoted to the slider task or the video task. This choice task resembled the choice task used in Experiment 1.

Participants were shown a "machine" that could create slider or video tasks (Figure 5c). The machine consisted of a black-and-gold "best" spinner and a panel of possible new spinners. Participants selected either "run best spinner" or "run new spinner". If the participant selected "run best spinner", the spinner would visually rotate on the screen. If it landed on gold, the machine created a video task; if it landed on black, the machine created a slider task.

If the participant selected "run new spinner", the new spinners in the panel were covered up and randomly shuffled. The participant then clicked one of the gray squares, revealing the new spinner underneath. As was explained to the particpants, and was visually apparent, one third of the possible new spinners are completely black, while the remaining two thirds range from 5% to 100% gold, in even increments of 5%.

After revealing a new spinner, it was spun, producing a video or slider task in the same manner as the "best spinner". As in Experiment 1, if the new spinner selected had a higher proportion gold than the best spinner, it would replace the best spinner for future choices.

Participants were also informed that after every work period there was a one in six chance that the machine would reset itself. In fact, the experiment was designed so that there was exactly one reset in every set of 6 trials (i.e., trials 1–6, 7–12, etc.). When the machine reset, the "best spinner" was set to a new starting value. The starting values following resets (including the intial starting value) were {20%, 25%, ... 55%, 60%}, randomly ordered.

Immediate and delayed conditions.

Participants were pseudo-randomly assigned to one of two conditions. In the Immediate condition, participants completed the task produced by a choice in the work period immediately following the choice. In the Delayed condition, participants completed the task produced by a choice after eight intervening work periods had passed, which was about five minutes after making the choice. This means that participants in the Delayed condition began making choices during the initial eight slider task work periods, in order to have outcomes determined when they reached the ninth and later work periods.

To make this delay intuitive, participants were shown a work queue at the bottom of the screen that contained eight tasks (see the bottom of Figure 5c). In the Delayed condition, upon making a choice a new slider or video task icon was added to the right of the queue, and then the leftmost task on the queue was performed and removed. In the Immediate condition, participants were still shown the cue, but upon adding an icon to the right of the queue that outcome was performed immediately. This means that in the Immediate condition the queue acted simply as a history of the past eight outcomes.

Post-task questions.

Following the final work period, participants were asked to rate their enjoyment of the slider task and of the video task on a scale from 1 to 7, where 1 indicated extremely unenjoyable and 7 indicated extremely enjoyable.

Results

Our primary hypothesis was that participants in the delayed-outcome condition would take more exploratory actions (that is, choose a new spinner more often) than those in the immediate-outcome condition. A secondary hypothesis was that this change would be moderated by participants' scores on the Barratt Impulsivity Scale. Specifically, we expected that highly impulsive participants would explore less and show a bigger difference in exploration between the delayed and immediate conditions.

We tested these predictions via hierarchical Bayesian logistic regression on participant choices. All aspects of this analyses were preregistered prior to data collection. We included predictors for the value of the current spinner, the participant's BIS score, the participant's condition, and interactions between condition and current spinner value and condition and BIS score. Condition was coded as -1 for the immediate condition and 1 for the delayed condition; current spinner value and BIS score were rescaled to have zero mean and unit variance across participants. We assumed that individuals could vary in their overall tendency to explore (i.e., intercept) as well as their responsiveness to current spinner value (slope). Participants' individual-level parameters were assumed to be drawn from a t distribution with $df = 5$, making our population level estimates robust to potential outliers. The priors on the the population-level predictor coefficients, and on the standard deviation of the t distributions from which individual-level parameters were drawn, were (truncated) normal distributions with a mean of 0 and a standard deviation of 5.

The model posterior was estimated using the Stan modeling language (Carpenter et al., 2017). We ran four chains of Hamiltonian Monte Carlo sampling, with 1000 samples per chain, the first half of which were discarded as burn-in. We confirmed convergence using the $\hat{R}$ convergence criterion (A. Gelman et al., 2014). In the results below, we report 95% credible intervals (CIs) on model parameters of interest. An overview of the model posterior is displayed in Figure 6

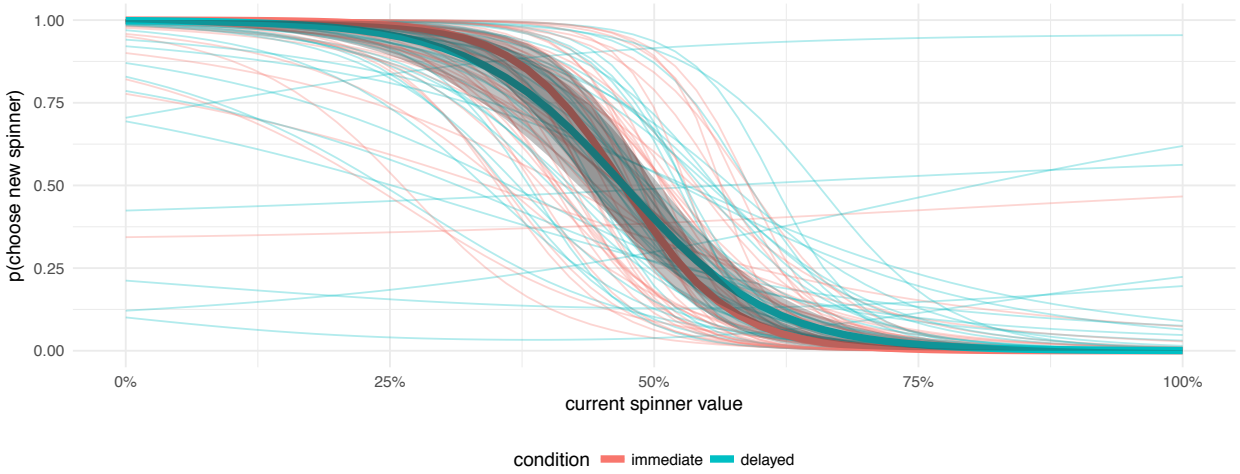

Figure 6: Model-based estimates of participants' probability of choosing a new spinner for different values of the current spinner in Experiment 2. Thick lines and shaded regions indicate the mean and 95% posterior interval for the population-level parameters, while the thin lines indicate the mean posterior parameters for each of the 100 individual participants. Participants in the delayed-outcome condition were no more likely to explore at a given current-spinner value than those in the immediate-outcome condition.

We found a strongly negative effect of current spinner value on participant's probability of selecting a new spinner, $CI = [-4.18, -3.08]$. This indicates that participants understood the general structure of the task, and explored (i.e., selected a new spinner) only when it was advantageous to do so. While the estimate was in the predicted direction, we found no clear effect of condition on the tendency to explore $CI = [-.11, .61]$. Additionally, there was no effect of BIS score on behavior $CI = [-.34, .36]$ and no interaction between BIS score and condition $CI = [-.21, .51]$. We did find a positive interaction between condition and current spinner

value, $CI = [.07, 1.06]$. This means that while people in the delayed condition were not more or less likely to explore in general, they were more likely to explore for high current spinner values, and less likely to explore for low current spinner values. In other words, they were less sensitive to the current value of the spinner.

Our preregistered analyses provided no support for our hypotheses. As an additional, exploratory analysis, we re-ran the Bayesian model replacing participants' BIS scores with their ratings difference between the slider task and the video task in post-experiment questionnaire. Over all, participants rated the video task as more enjoyable (6.37 out of 7 on average) than the slider task, (3.43 out of 7), $t(89) = 16.3, p < .001$. Our intuition was that participants who rated the video task much higher than the slider task may have felt a greater motivational pull to immediately watch a video instead of move slider, and may thus have been more susceptible to the delay manipulation. However, we found no main effect of ratings difference on exploration $CI = [-.32, .16]$, and no interaction between ratings difference and condition $CI = [-.33, .38]$. All other effects remained qualitatively the same.

Finally, we examined whether the lower sensitivity to current spinner value in the delay condition might indicate that a group of participants in that condition were responding near-randomly, possibly due to confusion with the task, and if this could affect our other results. We found that that the individual-level effect of current spinner value did not differ significantly from zero for 14 of 44 participants in the delayed condition, and only 2 of 46 participants in the immediate condition. To determine whether these near-random participants influenced our results, we re-ran our preregistered regression, including only the 30 participants with the highest-magnitude slopes in each condition from the initial analysis. We did not find that this new selection criterion affected our results. In particular, the credible interval for the main effect of delay still included zero, $CI = [-.08, .72]$.

Experiment 3

In both Experiments 1 and 2, we found no evidence that delaying rewards affected the degree of exploratory behavior, and thus no evidence that exploratory choice is influenced by present bias. Experiment 2 attempted to fix several flaws of Experiment 1: collecting data in person, making the consumption tasks more pleasant and aversive, increasing the reward delay, and simplifying the exploratory choice task. This may indicate that there is truly no effect of present bias on exploratory choice, but it remains possible that this null effect is due to a weakness in our experiment design.

The most apparent potential weakness is that the consumption tasks did not induce very much present bias, or that discounting of these stimuli occurred on a scale much longer than the delay of a few minutes used in our experiments. Our use of these stimuli was based on several past studies. Access to videos has been shown to induce present bias with a delay of around a minute (D. Navarick, 1998), and cessation of annoying noises can induce present bias with a delay of around ten seconds (Solnick et al., 1980). However, these studies were were small and differed from the current setting in important ways. For example Solnick et al. (1980) had participants make choices about noise cessation while solving math problems, which prevented them from focusing fully on the choice task.

It may be that the consumption tasks and setting we used did not, in fact induce present bias, which would mean that inducing a delay would have no predicted effect. Therefore, in Experiment 3 we conducted a simple follow-up experiment using the two consumption tasks to test whether people have a present bias for watching the video immediately, based on past experiments which studied time preferences for videos or video games (Millar & Navarick, 1984; D. Navarick, 1998).

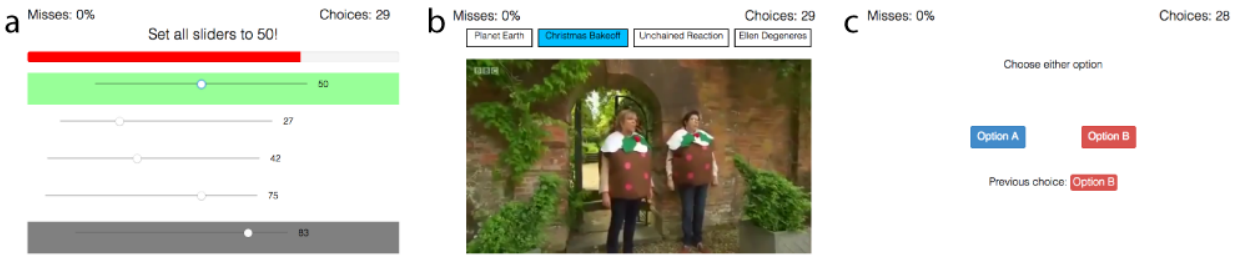

Figure 7: Examples of the Experiment 3 tasks. (a): an example of the slider task, in which the numeric timer has been replaced by a red timer bar. (b): an example of the video task. (c): an example of the decision-making task.

Methods

Participants.

Design and procedure.

As in Experiment 2, the Barratt Impulsiveness Scale was administered prior to completing the main task.

In the main task, shown in Figure 7c, participants were instructed that they would have to make a series of choices between two buttons. They were told that after selecting a button they would spend some amount of time performing a boring slider task and a fun video task, and that their choice could affect the amount of time spent on each task and the order of the tasks. They were also instructed that for their first two choices they would have to click first one button, then the other, to ensure that they had experienced both outcomes, and that occasionally the outcomes would change, at which point they would be instructed to try each of the two buttons again. On all other trials, they were told to select whichever button they preferred. Participants' previous choice was displayed at the bottom of the screen as a memory aid.

The slider and video tasks were very similar to the tasks used in Experiment 2 (Figure 7a/b). To prevent participants from explicitly measuring the amount of video and slider time following a choice, the timer showing how many seconds remaining in the consumption task was removed. For the slider task, it was replaced by horizontal red "progress bar" that steadily shrank over the course of the task, thereby indicating seconds remaining. For the video task, there was no indication of seconds remaining. In addition, instead of always lasting 30 seconds, the consumption tasks varied in length. For a slider task that lasted s seconds, there were $s/5 - 1$ sliders to complete.

After practicing the slider and video tasks, participants completed 30 trials of the choice task. This was divided into three groups of ten trials, each of which had a new pair of outcomes. The outcomes always lasted 90 seconds in total, and for each group there was always one button that produced the video task immediately, followed by the slider task, and one that produced the slider task immediately, followed by the videos. The reward amounts and reward orders of the three groups were as follows:

1. 30s videos/60s sliders vs. 60s sliders/30s videos
2. 35s videos/55s sliders vs. 65s sliders/25s videos
3. 25s videos/65s sliders vs. 55s sliders/35s videos

The ordering of the three pairs of outcomes was counterbalanced across participants, and the pairing of outcomes the left and right button was randomized. Absent discounting, participants should be indifferent between the two options in pair 1, and prefer the options with more video time in pairs 2 and 3. However, we predicted that while amount of video time would also matter, participants would display a bias towards selecting the option with the immediate video task.

As in Experiments 1 and 2, participants were asked to rate their enjoyment of the two consumption tasks following the experiment.

Results

Discussion

Reasons our study didn't work...

- Maybe our null result is wrong: other consumption reward literature has questionable statistics - maybe this really doesn't work that well? Need for new methods on how to quickly determine discount rates for consumption rewards.
- Maybe our null result is right: exploration uses other system than intertemporal choice. Curiosity, etc? Ways to belance exploration and exploitation without explicit consideration of the future. Fragility of time.

References

Abramson, L. Y., Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness Depression : A Theory-Based Subtype of Depression. *Psychological Review*, *96*(2), 358–372.

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: critique and reformulation. *Journal of Abnormal Psychology*, *87*(1), 49–74. http://doi.org/10.1037/0021-843X.87.1.49

Brown, A. L., Chua, Z. E., & Camerer, C. F. (2009). Learning and visceral temptations in dynamic saving experiments. *The Quarterly Journal of Economics*, (February), 197–231.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). http://doi.org/10.18637/jss.v076.i01

Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology*, *36*(5), 451–462. http://doi.org/10.1037/0022-3514.36.5.451

Evans, G. W., Gonnella, C., Marcynyszyn, L. a, Gentile, L., & Salpekar, N. (2005). The Role of Chaos in Poverty and Children's Socioemotional Adjustment. *Psychological Science*, *16*(7), 560–565.

Fang, C., Lee, J., & Schilling, M. a. (2010). Balancing Exploration and Exploitation Through Structural Design: The Isolation of Subgroups and Organizational Learning. *Organization Science*, *21*(3), 625–642. http://doi.org/10.1287/orsc.1090.0468

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351–401. http://doi.org/10.1257/002205102320161311

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2014). *Bayesian data analysis*. Retrieved from http://www.tandfonline.com/doi/full/10.1080/01621459.2014.963405

Gill, D., & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *The American Economic Review*, *102*(1), 469–503.

Gopher, D., Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to the training of complex skills. *Acta Psychologica*, *71*(1-3), 147–177. http://doi.org/10.1016/0001-6918(89)90007-3

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior*

Research Methods. http://doi.org/10.3758/s13428-015-0642-8

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Huys, Q. J. M., & Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, *113*(3), 314–328. http://doi.org/10.1016/j.cognition.2009.01.008

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., ... Prince, S. E. (1996). A Component Analysis of Cognitive-Behavioral Treatment for Depression. *Journal of Consulting and Clinical Psychology*, *64*(2), 295–304.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2016). Information Sampling Behavior With Explicit Sampling Costs. *Decision*.

Kirby, K. N., & Herrnstein, R. (1995). Preference Reversals Due To Myopic Discounting of Delayed Reward. *Psychological Science*, *6*(2), 83–89. http://doi.org/10.1111/j.1467-9280.1995.tb00311.x

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, *112*(2), 443–447. http://doi.org/10.1162/003355397555253

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: on the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392. http://doi.org/10.1037/a0023010

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, *14*, 95–112. http://doi.org/10.1002/smj.4250141009

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71–87.

Mcclure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time Discounting for Primary Rewards., *27*(21), 5796–5804. http://doi.org/10.1523/JNEUROSCI.4246-06.2007

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Fiedler, K. (2015). Unpacking the Exploration – Exploitation Tradeoff : A Synthesis of Human and Animal Literatures. *Decision*, *2*(3), 191–215.

Millar, A., & Navarick, D. J. (1984). Self-control and choice in humans: Effects of video game playing as a positive reinforcer. *Learning and Motivation*, *15*(2), 203–218. http://doi.org/10.1016/0023-9690(84)90030-4

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, *64*(3), 263–276.

Navarick, D. (1998). Impulsive choice in adults: How consistent are individual differences? *The Psychological Record*, *48*, 665–674.

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. http://doi.org/10.1016/j.cogpsych.2016.01.001

Otto, a. R., Markman, a. B., & Love, B. C. (2012). Taking More, Now: The Optimality of Impulsive Choice Hinges on Environment Structure. *Social Psychological and Personality Science*, *3*(2), 131–138. http://doi.org/10.1177/1948550611411311

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. http://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Rich, A. S., & Gureckis, T. M. (2017). Exploratory choice reflects the future value of information. *Decision*,

(in press).

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies*, *4*(2), 155–161.

Sang, K., Todd, P. M., & Goldstone, R. L. (2011). Learning near-optimal search in a minimal explore/exploit task. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 2800–2805.

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships: An experimental field test of the contact hypothesis: Research article. *Psychological Science*, *19*(7), 717–723. http://doi.org/10.1111/j.1467-9280.2008.02147.x

Solnick, J. V., Kannenberg, C. H., Eckerman, D. A., & Waller, M. B. (1980). An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation*, *11*(1), 61–77. http://doi.org/10.1016/0023-9690(80)90021-1

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit task. *Topics in Cognitive Science*, *7*. http://doi.org/10.1111/tops.12145

Stan Development Team. (2015). Stan: A C++ Library for Probability and Sampling, Version 2.7. Retrieved from http://mc-stan.org/

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge Univ Press.

Teodorescu, K., & Erev, I. (2014a). Learned Helplessness and Learned Prevalence: Exploring the Causal Relations Among Perceived Controllability, Reward Prevalence, and Exploration. *Psychological Science*, *25*(10), 1861–1869. http://doi.org/10.1177/0956797614543022

Teodorescu, K., & Erev, I. (2014b). On the Decision to Explore New Alternatives: The Coexistence of Under- and Over-exploration. *Journal of Behavioral Decision Making*, *27*, ‘. http://doi.org/10.1002/bdm

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

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*.

Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). How short- and long-run aspirations impact search and choice in decisions from experience. *Cognition*, *144*, 29–37. http://doi.org/10.1016/j.cognition.2015.07.006

Yechiam, E., Erev, I., & Gopher, D. (2001). On the potential value and limitations of emphasis change and other exploration-enhancing training methods. *Journal of Experimental Psychology. Applied*, *7*(4), 277–285. http://doi.org/10.1037/1076-898X.7.4.277

Zwick, R., Rapoport, A., Lo, A. K. C., & Muthukrishnan, a. V. (2003). Consumer Sequential Search: Not Enough or Too Much? *Marketing Science*, *22*(4), 503–519. http://doi.org/10.1287/mksc.22.4.503.24909