Delayed reward information is underweighted in reinforcement learning with dispersed feedback

Miruna Cotet[1,2]*, David Poensgen[3]*, Ian Krajbich[1,4+]

[1]The Ohio State University, Columbus, US
[2]Complexity Science Hub, Vienna, Austria
[3]Goethe University Frankfurt, Frankfurt, Germany
[4]University of California, Los Angeles, US
*These authors contributed equally.
+Corresponding author. Email: krajbich@ucla.edu

Abstract

Learning is fundamental to adaptive behavior. In the typical learning task, each action is associated with only one outcome, which could be immediate or delayed. However, actions often have multiple consequences that unfold over time. Here, we used behavioral and eye-tracking experiments to study how people learn when their choices yield both immediate and delayed reward information. Importantly, the rewards themselves were all delivered at the end of the study so there was no reason to weight immediate and delayed reward information differently. Instead, we found that our subjects overweighted immediate reward information. Moreover, this bias increased over the course of the experiment and was still present when learning from others' choices. The gaze data reveal mixed evidence that subjects looked more at immediate vs. delayed feedback, and across subjects, the relative dwell proportion did not predict the behavioral bias. Our results indicate that people prioritize not just immediate rewards, but immediate reward information. Unlike temporal discounting, this form of impatience is a clear mistake and leads to objectively worse outcomes.

Keywords: reinforcement learning, eye-tracking, memory, delay, attention

Introduction

Learning is fundamental to adaptive behavior [1–3]. To thrive, organisms must learn what things to pursue and what things to avoid. Reinforcement learning (RL) theory [4] is the dominant framework for understanding how organisms do this [5]. In the RL framework, learning involves updating reward expectations based on the difference between experienced rewards and prior expectations. With enough experience, one's reward expectation should converge to the true expected value [6,7].

In the typical RL task, each action is associated with only one reward, which could be immediate or delayed. However, outside the lab, actions often have multiple consequences that unfold over time [8]. In such cases it may be difficult for the organism to learn properly. For instance, credit-assignment is a well-known problem in RL where organisms sometimes misattribute rewards to the wrong actions [9,10]. This problem can occur when there are more actions than outcomes or in complex environments where there is ambiguity about which action(s) led to the reward. What remains unclear is whether people learn appropriately when there is no such ambiguity but there are more outcomes than actions, i.e., when a single action leads to both immediate and delayed rewards.

It is well known that organisms often discount the value of delayed rewards, meaning that they often prefer smaller immediate rewards over larger delayed rewards. This phenomenon is referred to as time discounting or delay discounting and usually thought of as a preference [11–15]. However, it is possible that at least some part of delay discounting is due to inaccurate learning. A difference in the processing of delayed relative to immediate feedback could generate discounting-like behavior that is based on learning frictions rather than preference.

There are many reasons why immediate and delayed reward feedback might have different effects on behavior. One reason is limited attention – people may prioritize attending to one kind of feedback over the other. Biased attention can lead to biased choices [16,17].

A second reason is agency – people learn better from things they choose compared to things they don't choose [18–21]. Delayed feedback may feel less agentic than immediate feedback. If agency-related boosts in reward prediction error only apply to the most recent choice, then one would expect a preference for options that yield larger immediate rewards.

A third reason is memory accessibility – people may find it easier to retrieve immediate feedback than delayed feedback at the time of choice, leading to a stronger influence of immediate vs. delayed feedback on choice. We know that people put more decision weight on information that is presented earlier [22], considered earlier [23,24], or retrieved from memory earlier [25,26]. If immediate feedback comes to mind more quickly than delayed feedback, for instance because of the speed of striatal associations, then this could produce a bias towards overweighting immediate feedback.

A fourth reason is an asymmetry in reinforcement learning – people may learn from delayed feedback at a slower rate than from immediate feedback. This possibility arises from research in neuroscience, indicating that immediate and delayed rewards are processed in distinct neural systems. The hippocampus has been shown to be selectively sensitive to delayed feedback, while the ventral striatum has been shown to be selectively sensitive to immediate feedback [27]. Moreover, amnesic patients with damage near the hippocampus are impaired at learning from delayed feedback but not immediate feedback, while patients with striatal dysfunction due to Parkinson's disease are impaired at learning from immediate feedback but not delayed feedback [28,29]. We also know that animals are impaired at learning when there are

longer delays between actions and outcomes [30,31]. Given this evidence that immediate and delayed rewards are processed independently, it seems plausible that these two types of feedback might not be properly integrated into a single, unbiased, total value. While we do not employ brain imaging or patient data in our study, we do explore the implications of two learning systems for behavior.

To study how people integrate dispersed feedback we use a reinforcement learning task in which decision-makers receive feedback about part of the total value immediately after their decision and the second, equally important part, one trial later. Thus, after every decision they need to learn about the option they just chose as well as the option they chose previously. Importantly, the rewards themselves were all delivered at the end of the study session so there was no reason to weight immediate and delayed rewards differently. In some trials, the stimulus with the larger total reward had a smaller immediate reward. Overweighting immediate rewards could lead to errors in these cases. We modeled learning in this task using an extension of a simple RL model [32], embedded in a dynamic choice model [33–36] testing whether our subjects learned from immediate rewards at a higher rate than from delayed rewards. Using eye-tracking, we also tested whether our subjects preferentially gazed at immediate rewards compared to delayed rewards. We correlated the individual learning rate parameters to the proportion of dwell time difference spent on either type of reward to test whether attention could explain any differences in learning rates. We also investigated whether any learning biases were linked to distorted declarative memories for those stimuli. Finally, we also tested whether any learning biases were related to working memory or intertemporal preferences.

To preview our results, we find that people are impaired at learning from delayed feedback, that this immediacy bias is somewhat reflected in aggregate but not individual gaze

measures, and that this behavioral bias may be linked to temporal discounting over much longer timescales. Our subjects put roughly twice as much weight on the immediate feedback as the delayed feedback when making their choices. In the aggregate our subjects tended to look more at the immediate feedback than the delayed feedback, though this tendency did not correlate with the behavioral immediacy bias. Interestingly, the behavioral immediacy bias was still present in a passive learning task where subjects observed others making the first 63 decisions. Finally, subjects who showed a greater immediacy bias in the learning task were, in one experiment, more likely to choose a smaller-sooner than larger-later reward in a standard intertemporal choice task with delays on the order of months. In summary, we find evidence that people do not optimally integrate immediate and delayed feedback and that the bias to discount delayed feedback may be linked to a more general tendency to discount the future.

Results

226 subjects learned the values associated with 6 stimuli, represented by colors in Study 1 (Colors Task; N = 102) and patterns in Study 2 (Patterns Task; N = 124; Fig. 1). An additional 114 subjects completed Studies 3–4 (Patterns Task), reported in the supplementary material. Each stimulus generated a different number of points. Feedback for each choice was not given at once but split into two components: one shown directly after the choice, and one shown after the following trial's choice. Thus, each stimulus had an immediate and a delayed feedback component. Both were equally important in determining the subject's earnings at the end of the study, which were based on the total points earned (SI Methods). Each feedback component had a baseline value (from 2 to 11 points) to which we randomly added up to 4 points in each trial (Fig. 1).

To identify whether subjects would overvalue options with higher immediate than delayed feedback, some stimuli had descending feedback (high immediate and low delayed) while others had ascending feedback (low immediate and high delayed). Each study had two groups of subjects and each group had stimuli that alternated between ascending and descending as the stimuli increased in total reward (Fig. 1).

There were 105 binary choice trials in each study, separated into blocks of 21 trials with every combination of stimuli. We excluded subjects whose accuracy on trials with both options ascending or descending was below 60 percent. We excluded trials in which the two options were identical as there was no (in)correct answer on such trials. We also excluded the first block of trials (21 trials) since subjects had very little information about the stimuli during that block. Finally, we also excluded trials in which subjects did not choose within the time limit (see Methods).

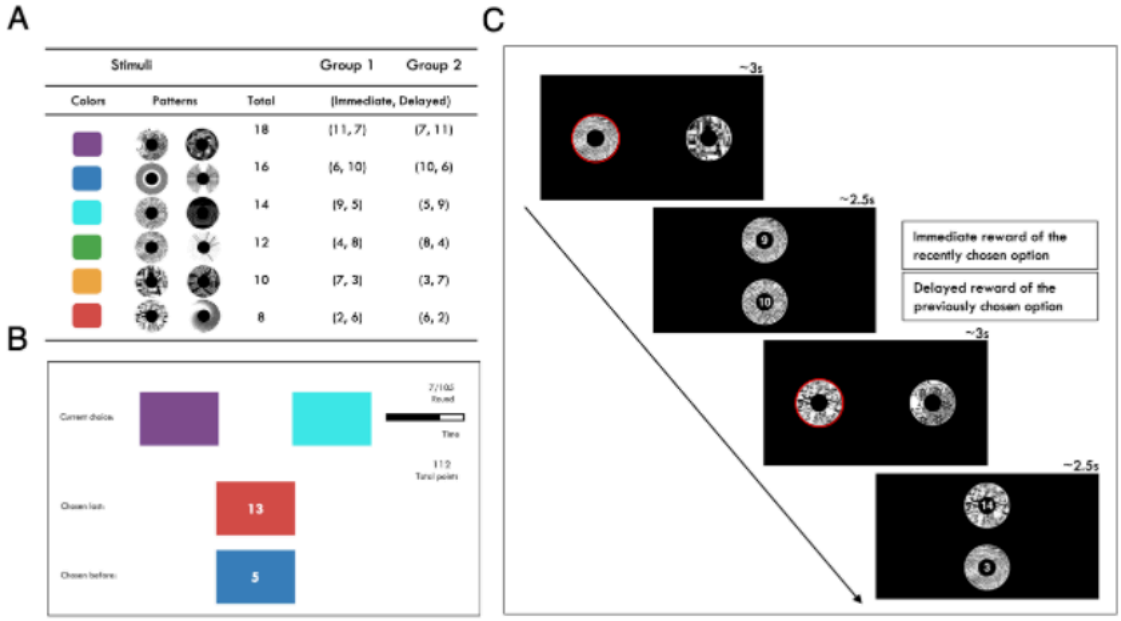


Figure 1. Experimental design. (A) For each subject/block there were six unique stimuli, colors in Study 1 or patterns in Study 2. Each stimulus had a total reward value ranging from 8 to 18, in steps of 2. For each total reward value there was an ascending version and a descending version based on whether the immediate feedback was smaller or larger than the delayed feedback, respectively. The two feedback components were set by dividing the total reward in half, adding 2 for the larger component, and subtracting 2 for the smaller. For each total reward value, Group 1 saw the descending or ascending version and Group 2 saw the opposite. A small random number was added to each underlying feedback component to make learning more difficult. Assignment of colors and patterns to total rewards was randomized by subject. (B) Study 1 screenshot. Each trial, subjects chose between two colored rectangles. Below the current options were the immediate feedback from the previous choice and the delayed feedback from the choice before that. Subjects also saw the trial number, the time within the trial, and the total accumulated points. After each choice, the boxes slid down the screen, feedback was presented, and then new options were shown. (C) Study 2 timeline. Each trial, subjects chose between two patterned discs. After each choice, subjects saw the immediate feedback in the center of their chosen disc. Below (or above) that, they saw the delayed feedback in the center of the chosen disc from the previous trial.

Behavior

To preview the behavioral results, we find that subjects displayed a learning bias in favor of immediate feedback. Subjects were more likely to choose a descending option than an ascending option, were more likely to make an error when the ascending option was the correct choice, and put more weight on the immediate feedback than the delayed feedback. Moreover, this bias increased over the course of the experiment.

For the following analyses, we have 87 subjects for Study 1 (43 in Group 1, 44 in Group 2) and 90 subjects for Study 2 (47 in Group 1, 43 in Group 2) after exclusions. We also excluded 1 trial in Study 1 and 61 trials in Study 2 due to subjects missing the time limit.

Looking at subjects' choices, we found that they put more weight on immediate feedback compared to delayed feedback. For the same total reward, the option with the higher immediate feedback was more likely to be chosen than its counterpart with the lower immediate feedback (Fig. 2A,B). Incongruent trials (worse option descending, better option ascending) had lower accuracy rates than congruent trials (worse option ascending, better option descending) as evidenced by a paired t-test on the average accuracy rates at the subject level (Study 1: congruent $M = 0.88$, incongruent $M = 0.53$, $t(86) = -9.28$, $95\%CI = [-0.42,-0.27]$, $p < 10^{-13}$; Study 2: congruent $M = 0.77$, incongruent $M = 0.62$, $t(89) = -4.34$, $95\%CI = [-0.21,-0.08]$, $p < 10^{-4}$).

We confirmed these results using regressions of *Choose Left* on differences in the average immediate, delayed, and total rewards between the left and right options, as well as whether the options were ascending or descending. We used mixed-effects regressions with random intercepts and slopes at the subject level. All continuous variables were z-scored. For each trial we calculated the relevant average feedback seen by the subject up to that point in the experiment. Choosing the left option was more likely when it was descending as opposed to

ascending (Study 1: $p < 10^{-13}$; Study 2: $p < .017$; Fig. 2D, Table S1) or when the right option was ascending rather than descending (Study 1: $p < 10^{-14}$; Study 2: $p < 10^{-4}$; Fig. 2D, Table S1), controlling for the difference in total reward. Subjects put larger weights on immediate than delayed feedback. When regressing choice on the immediate and delayed feedback differences, the weight on immediate was higher than on delayed by a factor of 2.4 in Study 1 and 1.4 in Study 2. (Study 1: $\beta_{Immediate} = 2.18$, $95\%CI = [1.97, 2.40]$, $\beta_{Delayed} = 0.90$ $[0.71, 1.10]$; Study 2: $\beta_{Immediate} = 1.54$ $[1.33, 1.75]$, $\beta_{Delayed} = 1.11$ $[0.91, 1.31]$; Fig. 2C, Table S2) A Likelihood Ratio Test comparing the immediate and delayed coefficients revealed significant differences (Study 1: $\chi^2(4, N = 87) = 860.74$, $p < 10^{-15}$; Study 2: $\chi^2(4, N = 90) = 350.1$, $p < 10^{-15}$).

The bias to overweight immediate feedback over delayed did not decrease as the experiment progressed; instead, it increased (Fig. 3). Building on the previous regression, we included interaction effects between trial number and the immediate/delayed feedback. The interaction of trial number and immediate feedback was positive and significant (Study 1: $\beta_{Immediate:Trial} = 0.38$, $95\%CI = [0.28, 0.49]$, $p < 10^{-12}$; Study 2: $\beta_{Immediate:Trial} = 0.34$, $95\%CI = [0.26, 0.43]$, $p < 10^{-15}$; Table S2), while the coefficient for the interaction of trial number and delayed feedback was also significantly positive but smaller (Study 1: $\beta_{Delayed:Trial} = 0.12$, $95\%CI = [0.04, 0.20]$, $p = .003$; Study 2: $\beta_{Delayed:Trial} = 0.19$, $95\%CI = [0.11, 0.26]$, $p < 10^{-6}$; Table S2). A Likelihood Ratio Test comparing the immediate and delayed interaction coefficients revealed that the increase in the delayed coefficient over time was significantly smaller than the increase in the immediate coefficient over time (Study 1: $\chi^2(4, N = 87) = 860.74$, $p < 10^{-15}$; Study 2: $\chi^2(4, N = 90) = 350.1$, $p < 10^{-15}$). This indicates that the immediacy bias increased over the course of the experiment.

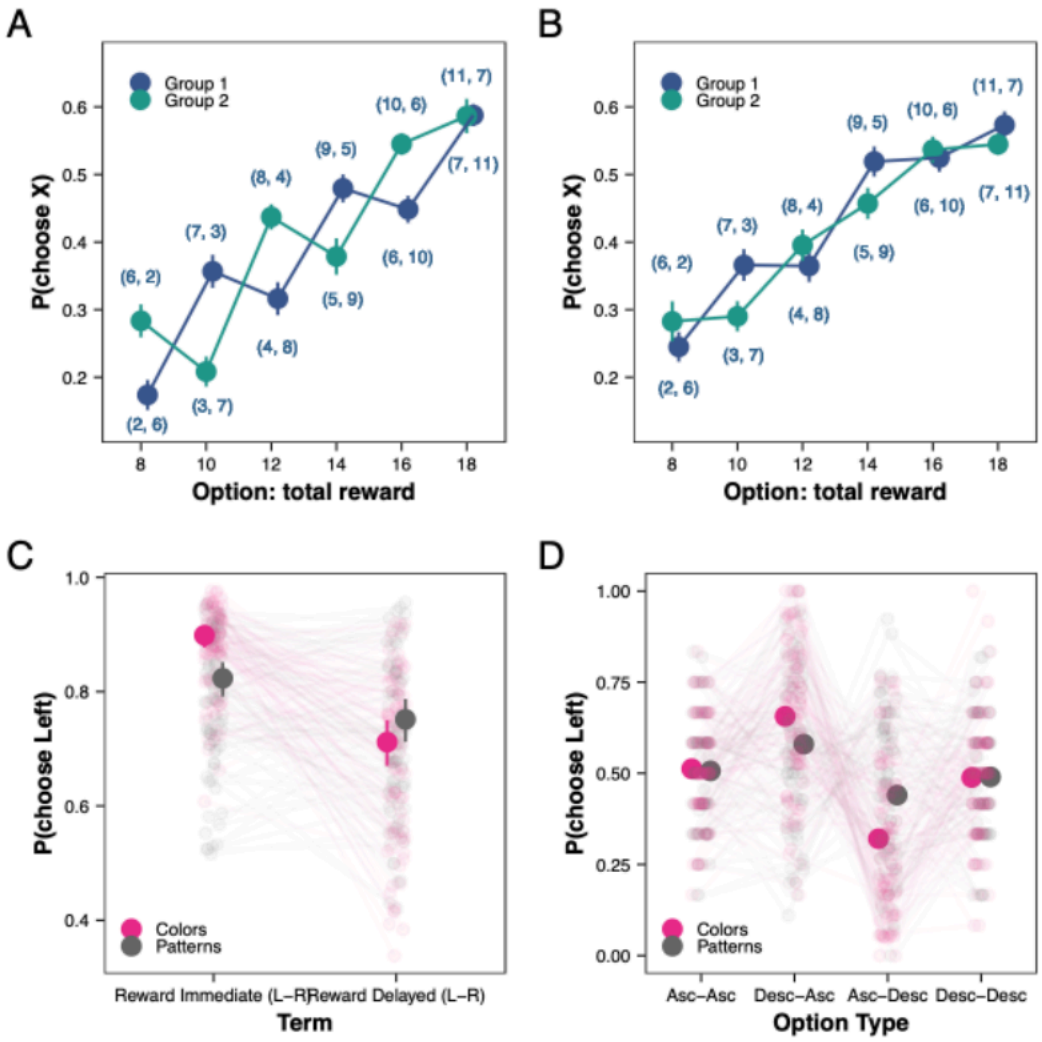


Figure 2. Choice behavior (A,B) Probability of choosing an option as a function of the option's total reward. (A) Study 1 (colors). (B) Study 2 (patterns). (C) Probability of choosing the left option as a function of the difference in the experienced immediate and delayed feedback, based on a mixed-effects logistic regression. Dots represent subject level effects and bars represent standard errors of the fixed effects. (D) Probability of choosing the left option as a function of whether the left and right options were descending or ascending. Dots represent subject level averages and bars represent standard errors across subjects.

The learning bias was also evident in subjects' response times (RT). We know from past work that people make faster choices when there is a larger absolute value difference (|VD|) between their options or a larger overall (summed) value (OV) of their options (for appealing options)[37–40]. Since subjects' choices were more influenced by immediate than delayed feedback, we expected to analogously see more influence of immediate than delayed feedback on RT. This was indeed the case. Immediate feedback had a larger effect on RT than delayed for both |VD| and OV. We regressed log(RT) on |VD| and OV, either in total rewards, or separated into immediate and delayed feedback. Larger total |VD| and larger total OV decreased RT (Study 1: $\beta_{|VD|} = -0.05$ $[-0.07, -0.04]$, $p < 10^{-13}$, $\beta_{OV} = -0.06$ $[-0.08, -0.05]$, $p < 10^{-12}$; Study 2: $\beta_{|VD|} = -0.04$ $[-0.05, -0.03]$, $p < 10^{-9}$, $\beta_{OV} = -0.05$ $[-0.06, -0.04]$, $p < 10^{-10}$; Figs. 4A,D & S14; Table S3). However, when we separated reward into immediate and delayed components, we found that immediate but not delayed |VD| significantly decreased RT (Study 1: $\beta_{VD\_Immediate} = -0.04$ $[-0.06, -0.03]$, $p < 10^{-9}$; $\beta_{VD\_Delayed} = -0.002$ $[-0.01, 0.01]$, $p = .786$; Study 2: $\beta_{VD\_Immediate} = -0.02$ $[-0.03, -0.01]$, $p < 10^{-3}$; $\beta_{VD\_Delayed} = -0.01$ $[-0.02, 0.00]$, $p = .056$; Figs. 4B,C & S14, Table S4). The difference between immediate and delayed effects was significant (Study 1: $\chi^2(6, N = 87) = 114.19$, $p < 10^{-15}$; Study 2: $\chi^2(6, N = 90) = 82.67$, $p < 10^{-14}$). We also found that immediate OV had a larger effect on RT than delayed OV (Study 1: $\beta_{OV\,Immediate} = -0.07$ $[-0.09, -0.05]$, $p < 10^{-12}$; $\beta_{OV\,Delayed} = -0.03$ $[-0.04, -0.01]$, $p < 10^{-4}$; Study 2: $\beta_{OV\,Immediate} = -0.04$ $[-0.05, -0.03]$, $p < 10^{-8}$; $\beta_{OV\,Delayed} = -0.03$ $[-0.04, -0.02]$, $p < 10^{-5}$; Difference: Study 1: $\chi^2(6, N = 87) = 95.29$, $p < 10^{-15}$; Study 2: $\chi^2(6, N = 90) = 52.07$, $p < 10^{-8}$; Figs. 4C,F & S14, Table S4).

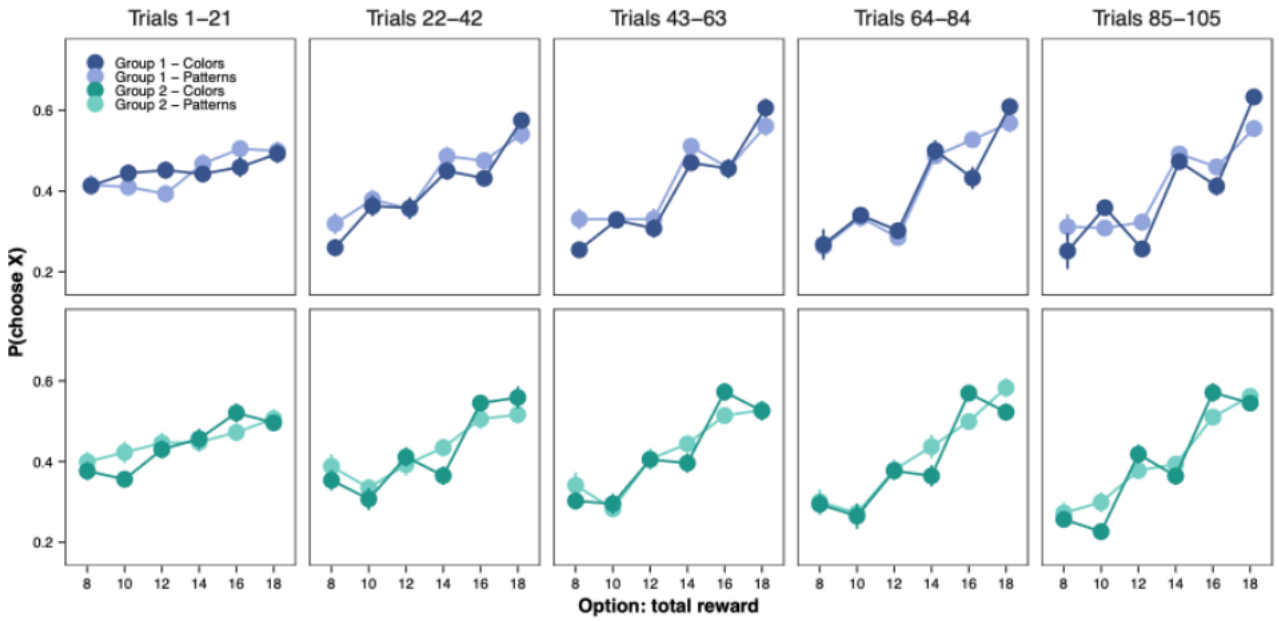


Figure 3. Behavioral bias over time. Probability of choosing a stimulus given it is in the choice set as a function of the total reward of the stimulus. The graphs show both the learning of total values (average slope) and the immediacy bias (zig-zag pattern) across trials. Though visually subtle, statistical analyses reveal that the immediacy bias increases across trials (Table S2). Dots represent subject level averages and bars represent standard errors across subjects.

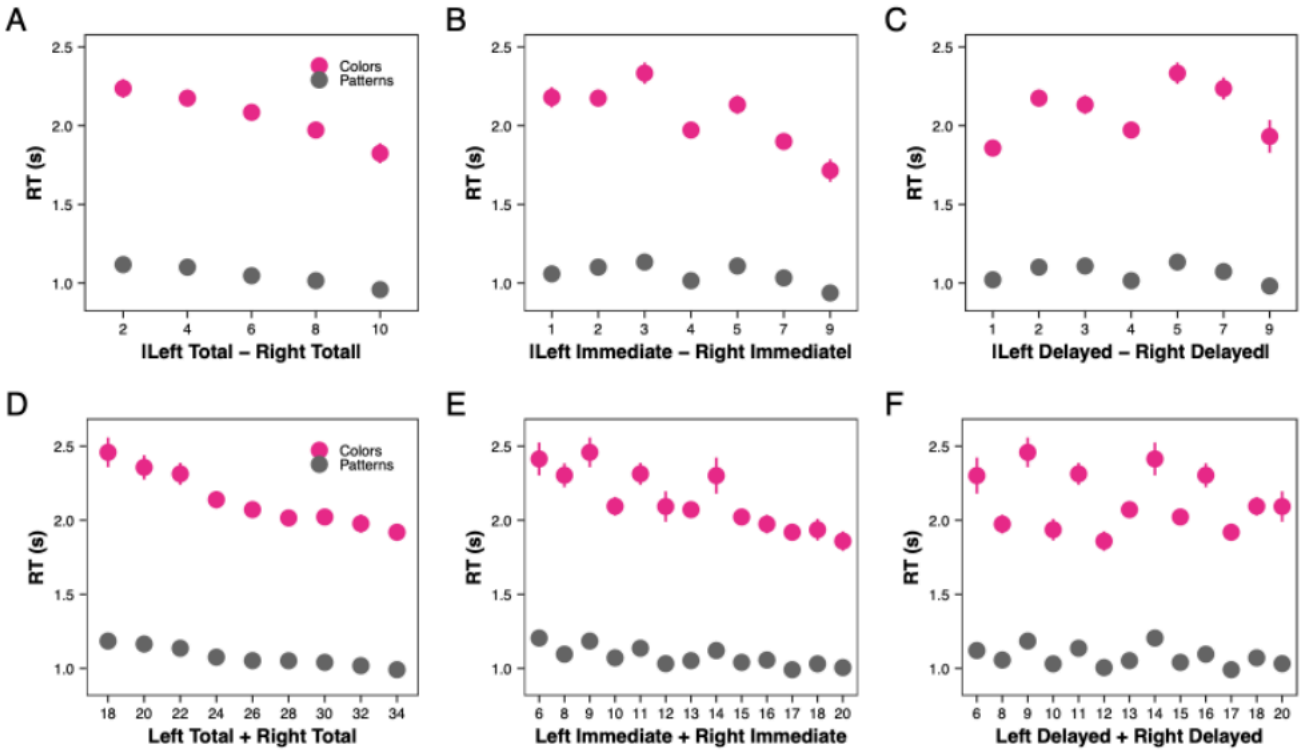


Figure 4. Value difference and overall value effects on response time (RT). (A,B,C) Absolute value difference (|VD|) effects on RT for each study. (A) RT decreases with total |VD|. (B) RT decreases with immediate |VD|. (C) RT does not decrease with delayed |VD|. (D,E,F) Overall value (OV) effects on RT for each study. (D) RT decreases with total OV. (E) RT decreases with immediate OV. (F) RT weakly decreases with delayed OV. Dots and bars represent mean and standard errors across subjects.

Computational Model

We sought to capture the relative effect of immediate and delayed rewards using a formal learning model. To do so, we employed an RL model embedded in a drift diffusion model (DDM), but allowed different learning rates for immediate and delayed rewards. We focus on this differential-learning model because our post-task data (described below) indicate an impairment during learning rather than decision-making, but later we also account for possible biases during choice.

We modeled 176 subjects, 87 subjects from Study 1 and 89 subjects from Study 2. One subject was excluded from Study 2's modeling analyses due to too few valid trials.

We fitted a reinforcement learning drift diffusion model (RLDDM) [33–35]. Given that each stimulus (s) was associated with two rewards, one immediate and one delayed (r^I,r^D), the total value for a stimulus was the sum of the predicted values for the immediate and delayed rewards (V). The two values generated two different prediction errors, one for the immediate reward and one for the delayed reward on each trial (k). The learning rates (α^I,α^D) controlled how much the values were updated after each reward. In order to determine if learning occurred at different rates for the immediate and delayed rewards, we also fitted a restricted model in which the learning rate was the same for both rewards ($\alpha^I = \alpha^D = \alpha$).

Consistent with the model-free analyses, we found higher learning rates for immediate rewards than for delayed rewards. This was true for both studies (Study 1: $M_{Immediate} = 0.27$, $M_{Delayed} = 0.14$, $t(86) = 6.97$, $95\%CI = [0.09, 0.16]$, $p < 10^{-9}$; Study 2: $M_{Immediate} = 0.23$, $M_{Delayed} = 0.18$, $t(88) = 2.75$, $95\%CI = [0.01, 0.09]$, $p = .007$; Fig. 5C). According to subject-level WAIC, the model with differential learning provides a better fit for 48% of our subjects while the model with the same learning rate for both types of rewards fit 52% of our subjects better. Thus, about half of our subjects were better described by two different learning rates.

The differential learning model fits the data well in an absolute sense. It reproduces the learning bias observed in the data; the option with the higher immediate reward was more likely to be chosen than the option with the higher delayed reward, for the same total reward level (Fig. S11). The model also shows increased accuracy across trials (Fig. 5A). The model does a good job of predicting accuracy for each type of choice trial (Fig. 5B) and the correlation between subject-level predicted accuracy and observed accuracy was high ($r(174) = .67$, $p < 10^{-15}$; Fig.

S12). The model predicts behavior more accurately on congruent and incongruent trials compared to when both options were ascending or descending (Fig. 5B).

In addition to the differential learning rates, it is also possible that people weight immediate and delayed rewards differently at the time of choice. This could occur either due to unequal attribute weighting [41] or due to discounting delayed feedback, relative to immediate feedback, during learning; these two mechanisms are mathematically equivalent assuming initial values are zero. To account for this possibility, we additionally fit models with different weights on immediate and delayed values, either with or without the different learning rates.

The model with equal learning rates and equal decision weights fit 32% of the participants best. Of the remaining participants, a majority (41%) were best fit by models that included different learning rates (best fits: unequal learning rates (14%), unequal decision weights (28%), or both (27%)). When considering data from all studies (87 for Study 1, 89 for Study 2, 40 for Study 3, 47 for Study 4), again the model with equal learning rates and equal decision weights fit 1/3 (33%) of the participants best. Of the remaining participants, a majority (53%) were best fit by models that included different learning rates (best fits: unequal learning rates (13%), unequal decision weights (31%), or both (22%)) (Table 1).

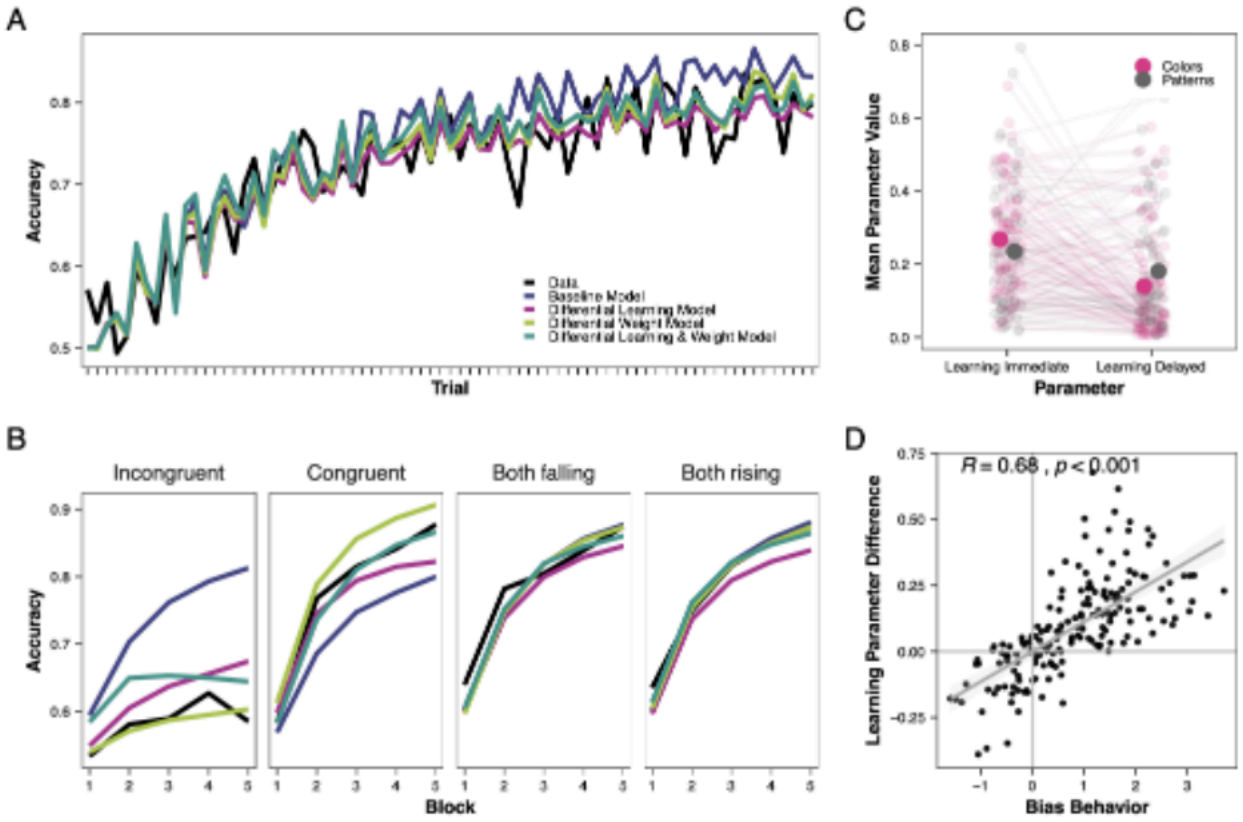


Figure 5. Model parameters and fits. (A,B) Average choice accuracy in the data and in the model simulations using the mean posterior values across trials and subjects. (A) Experiment level. (B) Block level for each type of trial. Incongruent choice sets: worse option descending, better option ascending, Congruent choice sets: worse option ascending, better option descending (C) Means and standard errors of the posteriors for the immediate and delayed learning rates. Dots represent subject-level learning rates for each task. (D) Correlation between behavioral bias and differential learning rate model bias. Behavioral bias is measured as the difference between the immediate reward coefficient and delayed reward coefficient in the regression of choice on rewards. Learning parameter difference is measured as the difference in the learning rates between the immediate and delayed reward.

	N	P	Mean (WAIC - WAIC Best)	Sum WAIC
Differential Weight Model	82	0.31	1.24	61100
Differential Learning and Weight Model	59	0.22	1.58	61198
Differential Learning Model	34	0.13	1.88	61306
Baseline Model	88	0.33	4.79	61715

Table 1. Model comparison results using WAIC. Total sample size from Studies 1-4 was 263 subjects. N and P are the number of subjects and proportion of subjects best fitted by each model. Mean (WAIC – WAIC Best) is the mean increase in WAIC for all the models that were not the best model. Sum WAIC is the sum of all WAIC across subjects.

Eye-tracking

As mentioned in the introduction, attention is one possible explanation for the differential impact of immediate and delayed feedback. To address this, we collected eye-tracking data in Study 2 (and Studies 3–4 in the supplementary material) while subjects saw feedback and made their choices. We sought to test whether subjects allocated more gaze to the immediate or delayed feedback, and whether the relative fraction of dwell time on the two feedback components predicted the learning bias. To preview the results, we observed a tendency to dwell longer on the immediate feedback compared to the delayed feedback, but this gaze bias was not correlated with the behavioral bias.

For these analyses, we used 75 subjects, after excluding 15 subjects who did not pass at least three out of four calibration checks throughout the experiment. We did not exclude any trials.

On the whole, subjects tended to look more at the immediate feedback compared to the delayed one (Fig. 6C). In a mixed-effects regression of relative dwell proportion on immediate vs. delayed, there was a marginal bias in favor of the immediate feedback when controlling for the size and type (ascending vs. descending) of reward ($\beta = 0.05$, 95% $CI = [0.00, 0.11]$, $p = .070$; Bayesian model: posterior mean $\beta = 0.05$, 95% CrI [–0.005, 0.11], $\Pr(\beta > 0) = .96$; Table S6),

and a significant bias when controlling for the predicted values and prediction errors (β = 0.09, 95%*CI* = [0.03,0.16], p = .007; Bayesian model: posterior mean β = 0.09, 95% *CrI* [0.02, 0.17], Pr(β > 0) = .996; Table S7). Using subject-level regressions, we found that 21 out of 75 subjects had a significant gaze bias towards the immediate feedback and 14 out of 75 had a significant gaze bias towards the delayed feedback.

Subjects were not more likely to fixate first on the immediate compared to the delayed feedback. In a mixed-effects logistic regression of first fixation location (immediate vs. delayed) there was no significant bias towards the immediate feedback, neither when controlling for the size and type of reward (β = 0.09, 95%*CI* = [−0.12,0.31], p = .379; Bayesian model: posterior mean β = 0.10, 95% *CrI* [–0.11, 0.32], Pr(β > 0) = .83; Table S6) nor when controlling for the predicted values and prediction errors (β = 0.21, 95%*CI* = [−0.07,0.49], p = .138; Bayesian model: posterior mean β = 0.21, 95% *CrI* [–0.07, 0.49], Pr(β > 0) = .93; Table S7). There was considerable heterogeneity across subjects in first fixation location. Using subject-level logistic regressions, we found that 12 out of 75 subjects had a significant first-fixation bias towards the immediate feedback and 11 out of 75 subjects had a significant first-fixation bias towards the delayed feedback.

Surprisingly, gaze biases did not positively correlate with choice biases. This was true both when using the behavioral bias calculated from the regressions (Dwell Proportion: Spearman's $\rho(73) = -.13$, p = .27; Bayesian analysis ρ = -.12, 95% *CrI* [-0.34, 0.10]; First Fixation: Spearman's $\rho(73) = -.25$, p = .028; Bayesian analysis ρ = -.24, 95% *CrI* [-0.44, -0.02]; Fig. S10) and when using the learning rates from the RL model (Dwell Proportion: Spearman's $\rho(73) = -.24$, p = .04; Bayesian analysis ρ = -.19, 95% *CrI* [-0.40, 0.03]; First Fixation:

Spearman's $\rho(73) = -.23$, $p = .05$; Bayesian analysis ρ = -.23, 95% *CrI* [-0.43, -0.01]; Fig. 6A,B). If anything, the correlation was in the opposite direction to what we expected.

However, we did replicate the general finding that options that receive more gaze during the choice phase are more likely to be chosen [16]. A regression of *Choose Left* on the dwell-time difference between left and right options revealed a positive and significant effect controlling for immediate and delayed feedback differences ($\beta = 0.43$, 95%*CI* = [0.30,0.56], $z = 6.36$, $p < 10^{-9}$; Table S8).

Studies 3-4 (Study 3 online, Study 4 in lab) yielded slightly different findings. In these studies, the correlations between gaze bias to the immediate feedback and choice bias were positive (rather than negative as in Study 2), but were only significant or marginal in Study 3 when looking at the choice bias from the regressions, and marginal in Study 4 when looking at the learning rates. There was also no significant overall dwell bias towards the immediate feedback in either study, though there was a significant first-fixation bias towards the immediate feedback in Study 4 (Figs. S10, S20–S21). Overall, these additional studies suggest that there may be a link between gaze during the feedback phase and choice bias, but that it is likely a weak relationship.

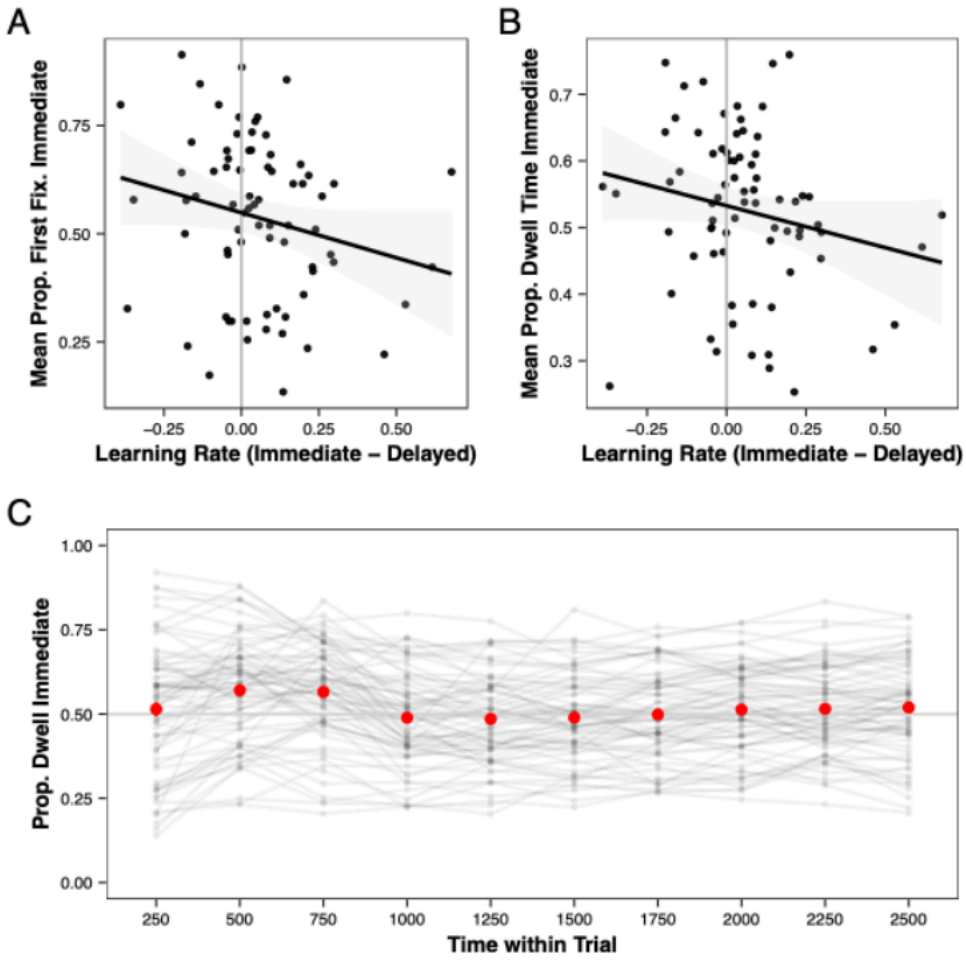


Figure 6. The link between gaze and behavior in Study 2. (A) The correlation between the difference in learning rate (immediate vs. delayed) and the proportion of first fixations to the immediate feedback (Spearman's $\rho(73) = -.23$, $p = .05$). (B) The correlation between the difference in learning rate (immediate vs. delayed) and the mean dwell proportion to immediate feedback across trials (Spearman's $\rho(73) = -.24$, $p = .04$). (C) The mean dwell proportion to immediate feedback within a trial for 250 ms time bins across the trial. Black dots represent each subject within a time bin and red bars are standard errors across subjects. (A,B) Dots represent each subject. The gray bands represent 95% CI.

Passive learning task

As described earlier, one possible explanation for the observed immediacy bias is agency. People learn better from things that they choose than things they don't choose, and so it is possible that the immediacy bias is due to an increased sense that immediate vs. delayed feedback is due to participants' choices.

To investigate this possibility, we conducted an additional experiment with passive learning. N=57 participants completed this passive learning experiment, using the color paradigm from Study 1. Participants first observed 3 blocks of choices from a Study 1 participant. The feedback and timing of the feedback was the same as that of the matched partner, except that they could not see the foregone choice option on the choice screen. After observing 63 trials, participants then made 2 blocks (42 trials) of choices on their own.

Participants in the passive learning experiment still exhibited a substantial immediacy bias in their choices. When regressing choice on the immediate and delayed feedback differences, the weight on the immediate feedback was substantially higher than the weight on delayed feedback ($\beta_{Immediate}$ = 2.62, 95%CI = [2.11, 3.12], $p < 10^{-15}$, $\beta_{Delayed}$ = 1.48 [1.07, 1.88], $p < 10^{-12}$, Fig. 7, Table S2). While the immediacy bias was perhaps smaller than in Study 1 (depending on the analysis, see SI), passive learning clearly does not eliminate the effect (Fig. 7).

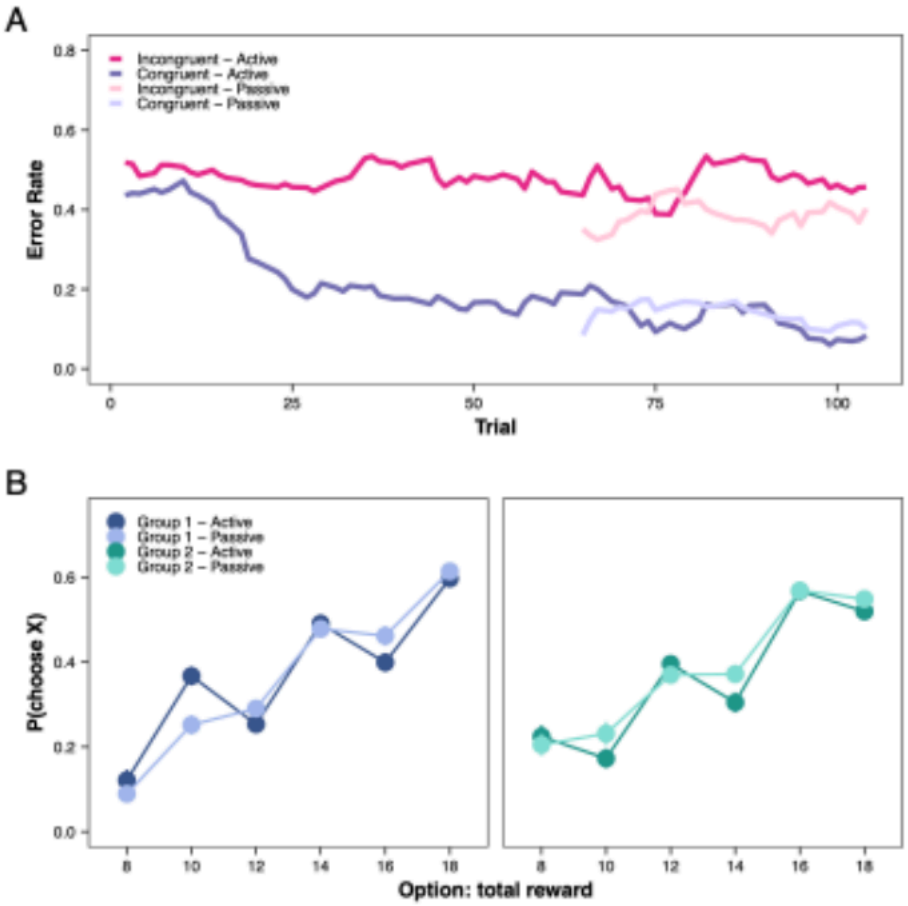


Figure 7. Passive learning experiment (a) Error rates and (b) choice probabilities for the passive learning experiment, compared to Study 1 (active learning matched participants) (a) split into congruent and incongruent trials, and (b) split into Group 1 and Group 2. Darker colors denote the active learning data while lighter colors denote the passive learning data.

Declarative Memory

At the end of Study 2 (and Study 4), we asked subjects to rank the stimuli in terms of total reward, to indicate whether each stimulus was ascending, descending or flat, and to estimate the average total reward for each stimulus. The results from these extra measures indicate that the behavioral immediacy bias is at least partly due to inaccurate learning, with biases in memory generally biased in the same direction as subjects' choices during the main task.

Being worse at ranking the stimuli was associated with lower overall accuracy (Study 2: β = 0.04, 95%CI = [0.03, 0.05], $p < 10^{-7}$, Study 4: β = 0.03, 95%CI = [0.02, 0.05], $p < 10^{-3}$; Table S9). Although ranking accuracy was not significantly associated with the immediacy bias (i.e., the relative size of the choice coefficients on immediate vs. delayed feedback) in Study 2, it was significantly positively associated with the immediacy bias in Study 4 (Study 2: β = –0.05, 95%CI = [-0.17, 0.08], p = .473, Study 4: β = 0.19, 95%CI = [0.03, -0.36], p = .02; Table S10). This indicates that correctly ranking the stimuli was associated with a higher, not lower, immediacy bias, at least for the in-lab study.

Having worse memory of whether a stimulus's immediate feedback was larger or smaller than its delayed feedback was associated with higher accuracy for incongruent choice sets (Study 2: β =- 0.07, 95%CI = [-0.12, -0.02], p = .006, Study 4: β = -0.11, 95%CI = [-0.18, -0.04], p = .004; Table S9). Moreover, these memory errors were associated with a lower immediacy bias (Study 2: β = -0.44, 95%CI = [-0.68, -0.21], $p < 10^{-3}$, Study 4: β =- 0.28, 95%CI = [-0.56, -0.01], p = .04; Table S10). In other words, better memory for the relative size of the immediate vs. delayed feedback was associated with a higher, not lower, immediacy bias in both studies.

In Study 4 we asked subjects to separately estimate the average for the immediate and delayed feedback components. Subjects made larger errors estimating points for delayed feedback than immediate feedback. The average absolute point error was 1.98 (SD = 0.51) for immediate feedback and 2.68 (SD = 0.73) for feedback rewards. The difference was significant (M = -0.7, $t(45)$ = -4.97, $p < 10^{-4}$). Errors in the estimation of the immediate feedback were associated with a lower immediacy bias (Study 4: β = -0.61, 95%CI = [-1.16, -0.05], p = .03; Table S10), but errors in the estimation of the delayed feedback were not associated with immediacy bias (Study 4: β = 0.08, 95%CI = [−0.42, 0.59], p = .74). In other words, better

448 memory of the immediate feedback components was associated with a higher, not lower,
449 immediacy bias.
450 Overall, better memory was associated with a higher, not lower, immediacy bias. This
451 argues against the notion that subjects exhibited immediacy bias due to inattention to the task.
452 On the other hand, biases in memory were correlated with biases in behavior.
453
454 ***Working Memory***
455 At the end of Study 1, subjects also completed a working-memory task – an n-back task
456 using either n=2 or n=3. We regressed the accuracy and d-prime measures for both n-back tasks
457 on the immediacy bias. For the 2-back task, accuracy was marginally negatively related to the
458 immediacy bias based on the regression approach, and significantly so based on the RL model
459 (Regression-based: $\beta = -2.85$ [−5.98,0.28], $p = .074$, RL-based: $\beta = -0.55$ [−1.02,−0.09], $p =$
460 .019; Table S11). The d-prime measure was also negatively related to the immediacy bias using
461 both measures (Regression-based: $\beta = -0.31$ [−0.60,−0.03], $p = .03$, RL-based: $\beta = -0.05$
462 [−0.10,−0.01], $p = .012$; Table S11). However, the association was not present for the 3-back
463 task (Regression-based: $\beta = -0.12$ [−0.54,0.30], $p = .556$, RL-based: $\beta = -0.02$ [−0.09,0.05], $p =$
464 .519; Table S11).
465 In Study 4, subjects instead completed a visual working memory task, namely a change
466 localization task. There was no relationship between subjects' performance on that task and their
467 immediacy bias (see supplementary material).
468
469 ***Discounting Preferences***

At the end of Study 1 (and Study 4), subjects completed a hypothetical survey to assess their intertemporal preferences. We then regressed the intertemporal indifference point for all three time-scales on the immediacy bias. The only significant association was for the decisions between today and one month. The stronger the behavioral bias, the higher the impatience, significantly for the regression-based analysis ($\beta = 2.04$, $95\%CI = [0.15, 3.92]$, $p = .034$; Table S12) and marginally for the RL-based analysis ($\beta = 11.04$ $[-1.69, 23.78]$, $p = .088$; Table S12). However, this was marginal or not significant for the other time-scales: today and 6 months (Regression-based: $\beta = 1.86$ $[-0.33, 4.05]$, $p = .096$, RL-based: $\beta = 8.03$ $[-6.79, 22.84]$, $p = .284$; Table S12) and 1 month and 6 months (Regression-based: $\beta = 1.53$ $[-0.54, 3.59]$, $p = .146$, RL-based: $\beta = 3.63$ $[-10.36, 17.632]$, $p = .607$; Table S12).

None of these relationships were significant in Study 4 (Table S13). This casts doubt on the relationship between intertemporal preferences and immediacy bias in our learning task.

Discussion

In this article we found that people sub-optimally overweight immediate reward feedback relative to delayed reward feedback. In our task, subjects merely had to add together two (small) numbers to determine the total reward from their choice. And yet, subjects put between 1.4 and 2.4 times as much weight on the immediate feedback compared to the delayed feedback, roughly equivalent to counting the immediate feedback twice. This resulted in 15 to 35% (Study 1: 35%, Study 2: 15%, Study 3: 17%, Study 4: 25%) increases in errors when the option with the larger immediate feedback was the wrong choice. Surprisingly, this immediacy bias only grew stronger as the experiment progressed. At the same time, we also observed an attentional bias towards immediate feedback – people spend about 53% of the time looking at the immediate feedback

493 (Studies 2 and 3: 52%, Study 4: 54%). However, this gaze bias generally does not correlate with
494 the choice bias. We also found mixed evidence that this immediacy bias may be linked to
495 impatience and working memory deficits.

496 Our analyses, which focused on both choice and RT, consistently revealed a stronger
497 influence of immediate feedback on subjects' choices. This behavioral bias favoring immediate
498 feedback was further evident in RT, with immediate feedback exerting a larger effect on RT via
499 both value difference and overall value. While RT significantly decreased in both the sum of and
500 the difference between the immediate feedback components of the two options, it showed
501 reduced or no such relationship with the delayed feedback, respectively. The relationship
502 between value-difference, overall value, and RT has been well documented [16,17,34,37–40,42–
503 48] and so its absence/reduction for delayed feedback lends further credence to the idea that
504 people prioritize the immediate feedback.

505 Overall, we find substantial support for an asymmetry in reinforcement learning whereby
506 people learn more slowly about delayed feedback than immediate feedback. We find mixed
507 support for other possible mechanisms – limited attention, agency, and memory accessibility. Our
508 eye-tracking data indicate that limited attention is unlikely to be the whole story. We find only
509 weak evidence that people fixate more on immediate vs. delayed feedback. Our passive-learning
510 data indicate that agency is unlikely to be the whole story. We still find a substantial immediacy
511 bias when participants were yoked through others' choices, though the bias did decrease
512 somewhat. Finally, our memory tests at the end of the experiments indicate that unequal
513 weighting of immediate vs. delayed feedback is unlikely to be the whole story. Our participants
514 were worse at recalling delayed vs. immediate feedback, and for a majority of them, their
515 behavior was better fit by a model that incorporated different learning rates. Thus, while

attention, agency, and memory accessibility might each marginally contribute to an immediacy bias, we conclude that the bulk of the bias is due to asymmetric learning.

We used an RL model embedded in a DDM to better understand and quantify the prioritization of immediate feedback. We incorporated distinct learning rates for immediate and delayed feedback, which provided a better fit for approximately half of the subjects. Immediate learning rates were significantly higher than delayed learning rates at the group level. While the two-learning-rate model provided generally good fits to the behavior, it isn't an entirely satisfying explanation for the immediacy bias because it predicts that the immediacy bias should shrink over time, while instead we find that the immediacy bias grows over time. Moreover, it under-predicts accuracy when both options are ascending or descending (particularly ascending). Future work will need to investigate other mechanisms by which this immediacy bias might occur.

Consistent with a bias in learning, during the feedback phase we observed some evidence for a greater focus on immediate feedback. In Study 2 subjects spent a greater proportion of time dwelling on immediate vs. delayed feedback, while in Study 4 subjects were more likely to look at the immediate feedback first. However, these attentional biases showed inconsistent correlations with the behavioral bias. In Study 2 the correlations were insignificantly negative, while in Studies 3-4 the correlations were positive but only sometimes significant. The eye-tracking results from Studies 2-3 do need to be interpreted with some caution as the data were collected using online webcam-based eye tracking. Collecting eye-tracking data online has many advantages but it does sacrifice some precision relative to in-lab eye-tracking systems [41,49,50]. Thus, we put more weight on our Study 4 results, which revealed a marginally positive correlation between dwell proportion on the immediate feedback and the learning rate advantage

for the immediate feedback. Overall, gaze may play a small role in the immediacy bias but it doesn't seem to fully explain the behavior.

Also consistent with a bias in learning, we found that subjects with a stronger immediacy bias in their choices were also better on the end-of-study memory test where they were asked to report whether each stimulus was ascending, descending, or flat. Moreover, their memory for the rankings or points for the stimuli were, if anything, positively correlated with the immediacy bias. These results indicate that the immediacy bias we observed was not simply due to a lack of attentiveness. Subjects displaying stronger immediacy bias were better at recalling the immediate feedback but not the delayed feedback, suggesting that the immediacy bias may be due to overweighting immediate feedback rather than underweighting delayed feedback.

The act of choosing the alternative that has the higher immediate reward instead of considering how current choices affect future rewards has been termed melioration [51–53]. Melioration at the expense of maximization has been demonstrated in many animal experiments [54,55] as well as with human subjects [56]. Previous research has used the Harvard game to show the tendency of melioration over maximization [57–62]. However, the task puts subjects in a highly complex learning environment in which any deviation from the optimal choice appears impatient. Even Bayesian agents with perfect memory often need thousands of trials to avoid melioration and arrive at the optimal solution. The reason lies in the opacity and complexity of that task. Our task explicitly addresses this criticism by being much simpler and fully transparent and allows us to cleanly attribute the observed bias to overweighting of immediate feedback [63].

Our results are consistent with research in neuroscience that has argued for different learning systems based on immediate and delayed feedback. The dopamine RL response to

562 delayed rewards is similar to rewards that are entirely unpredicted [64,65]. This indicates that the
563 midbrain-striatal system is not entirely able to learn from delayed rewards [66]. Therefore,
564 learning of delayed rewards might be supported by a different system [27]. There is also
565 evidence for distinct involvement of the striatum and hippocampus for different types of
566 feedback. This research argues that learning from immediate rewards is supported by an implicit
567 memory system associated with the ventral striatum (VS), while learning from delayed rewards
568 is supported by an explicit memory system associated with the hippocampus [27,28,66–68]. Both
569 neuroimaging and patient data support the dissociation between these two learning systems –
570 Parkinson's patients (with VS dysfunction) are impaired at learning from immediate but not
571 delayed feedback, while amnesic patients (with hippocampal dysfunction) are impaired at
572 learning from delayed but not immediate feedback [27–29]. There is some debate about whether
573 such evidence conclusively supports multiple systems [69]. It is also unclear whether this multi-
574 system account can explain our results. Unlike these studies, we re-displayed the chosen stimuli
575 when showing both immediate and delayed feedback. So, while the second feedback was
576 delayed relative to the action, it was immediate relative to the stimulus. In any case, future work
577 will need to examine the neural mechanisms underlying the immediacy bias presented here.
578 Our study contributes to the understanding of temporal discounting by providing an
579 alternative explanation for its prevalence and resistance to correction [11]. Unlike traditional
580 intertemporal choice paradigms, which focus on known options and delays, our research
581 highlights the role of learning in shaping impulsive behavior. If reinforcement loses efficacy with
582 delay, immediate consequences would receive undue weight relative to delayed ones. This would
583 lead to a discounting of delayed rewards, not because people do not like to wait, but because the
584 delayed rewards are harder to learn. Thus, the learning bias that our study demonstrates may be a

585 natural explanation for why impatience is so common, while the opposite, future bias, is so rare.
586 Our results offer some evidence in this direction by showing a correlation of behavior in the
587 learning task with typical measures of time discounting. Moreover, if time preferences are
588 influenced by experience rather than inherent traits, discount rates should vary across domains
589 within individuals. Empirical findings support this notion [70].

590 Our study was designed to rule out temporal discounting effects. First, our delays were on
591 the order of seconds rather than hours or days, as typical in intertemporal choice tasks.
592 Nonetheless, some studies have shown that temporal discounting can occur even with delays on
593 the order of seconds [64,71]. More importantly, reward discounting can not explain the bias
594 because all rewards were delivered at the end of the experiment. One might think that subjects
595 could have willingly foregone monetary reward in order to receive good feedback a few seconds
596 earlier. The high losses observed due to the bias make this implausible. In study 1 for example,
597 subjects on average lost €1.73 in incongruent and €0.64 in congruent trials, meaning that the
598 willingness to pay for good early feedback must have been at least €1.09 to explain the bias in
599 this manner ($t(86) = -6.75$, $p < 0.001$). Finally, the memory tasks indicate that subjects indeed
600 misestimated total rewards as reflected in their choices.

601 Cognitive load is another potential explanation for learning frictions. Our task required
602 subjects to process feedback from two different choices at the same time, one delayed and one
603 immediate. This could explain why subjects learn incompletely, but it doesn't explain why they
604 favor the immediate feedback – cognitive load should affect the immediate and delayed feedback
605 equally. Furthermore, once a subject has learned one feedback component (e.g., the immediate
606 one) they should shift their attention and learn the other, leading to a decrease in bias over time.
607 This is the opposite of what we observed. One possibility is that people might be holding in mind

608 the immediate feedback in order to add it to the delayed one. This might explain why the
609 immediate feedback seems to be double counted. Cognitive load has been shown to increase
610 dwell time [72] and this might be one reason why we don't observe an association between dwell
611 time and the immediacy bias. In any case, future work could study our task in a less constrained
612 setting where immediate and delayed feedback are given in isolation [27].

613 The prevalence of myopic behavior has significant implications across various domains.
614 In financial decision-making, it can lead to inadequate investments in retirement savings and
615 education, as well as overspending and debt accumulation. Myopic behavior is also associated
616 with health-related choices, where immediate gratification often takes precedence over long-term
617 well-being. Skill development and learning are affected by temporal discounting, as individuals
618 may prioritize immediate enjoyment over the sustained effort required for skill acquisition. This
619 can impact academic and professional performance, personal goals, and overall life satisfaction.
620 This bias could also be relevant in the realm of social media. The focus on immediate rewards,
621 coupled with algorithmic promotion of viral content, can influence users' sharing and liking
622 behaviors, potentially shaping online discourse and content consumption. Addressing and
623 mitigating the effects of delayed discounting may involve strategies such as enhancing the
624 visibility of future rewards [22] and emphasizing learning through observation.

625 In summary, we have shown that people show a bias to learn more from immediate
626 feedback than even slightly delayed feedback. This has major implications for how we evaluate
627 people's choices. As opposed to ascribing impatient behavior to an unwillingness to wait, we
628 argue that impatience may in part be due to delayed rewards receiving too little weight in
629 learning, perhaps due to inattention to delayed feedback. This may explain why impatience and
630 temporal inconsistency are such pervasive problems.

Materials and Methods

Ethics Statement

Study 1 was ruled as exempt by the joint Ethics Committee of the Faculty of Economics and Business Administration of Goethe University Frankfurt (GU) and the Gutenberg School of Management & Economics of the Faculty of Law, Management and Economics of Johannes Gutenberg University Mainz (JGU). Written informed consent was obtained prior to data collection.

Studies 2 & 3 were approved by the Behavioral and Social Sciences Institutional Review Board at the Ohio State University, protocol 2013B0583. Written informed consent was obtained prior to data collection.

Study 4 was ruled as exempt by the UCLA Office of the Human Research Protection Program. Written informed consent was obtained prior to data collection.

Experimental Paradigm

In Study 1, subjects had 10 seconds to make their decisions and lost 5 points if they ran out of time. In Study 2, subjects had 3 seconds and lost 5 points if they ran out of time. In Study 2, the feedback screen was presented for 2 seconds, and a randomly generated time interval between two and six seconds was added after each choice and feedback screen.

Study 1 was programmed using oTree [73]. Study 2 was programmed using jsPsych [74] and utilized webcam-based eye tracking using the Webgazer library [49] integrated in jsPsych [41].

At the end of each study, subjects completed a memory survey. We asked subjects to rank the stimuli in terms of total reward, to indicate whether each stimulus was ascending, descending or flat, and to estimate the average total reward for each stimulus (SI Methods). We only analyzed the memory surveys for Study 2, since the memory survey for Study 1 only asked the questions for a subset of the stimuli presented (SI Methods). For the ranking question we used Kendall tau distance between the true rank and the recalled rank. This measure counts the number of pairwise disagreements between two rankings. For the ascending vs. descending question, we computed the sum of the errors. If the subject answered that the immediate feedback was equal to the delayed feedback, we counted it as half an error. For the total reward question, we calculated the average error.

At the end of Study 1, subjects also completed a temporal discounting task and a working-memory n-back task [75].

In the temporal discounting task, subjects completed a hypothetical survey to assess their intertemporal preferences. In three series of questions, subjects were asked how they would decide between 100 euros today or x in one month; 100 euros today or x in six months; 100 euros in one month or x in six months. By increasing or decreasing x from question to question, we obtained indifference points in the range 100–132 euros for each time horizon.

In the n-back task, participants were presented with a sequence of stimuli, in this case numbers, one at a time. The task required participants to indicate whether the current stimulus matched the one presented n items back in the sequence. One sequence was a 2-back task and consisted of 48 stimuli of which 14 were target stimuli and one sequence was a 3-back task and consisted of 48 stimuli of which 16 were target stimuli.

We assessed performance on the n-back task using accuracy and d prime. Accuracy includes both correctly identified targets and correctly identified non-targets. D prime is a measure of sensitivity or discriminability used in signal detection theory [76] and is calculated as the difference in z-scored hit rates and false alarm rates. These measures were corrected to account for cases in which the hit rate equals 1 or false alarm rate equals 0. In order to correct the measures, we added 0.5 to the total number of hits and false alarms and added 1 to the total number of stimuli [77,78].

The 3-back task was not used for some subjects. In total, 87 subjects completed the 2-back task and 51 subjects completed the 3-back task. 1 subject was excluded from the analyses because their d prime measure was negative, indicating worse-than-chance behavior.

For Study 2 and 3, we excluded some participants based on their eye-tracking data. A short validation phase was presented after trials 21, 42, and 84. After 63 trials participants completed a new calibration and validation. If participants failed all 3 short validations, they were excluded. Each validation check consisted of 3 dots at different positions on the screen that participants had to fixate on. A participant failed the validation if they failed to fixate on any of the 3 dots, within the error margin.

Modeling

Models

For the differential learning model, the values for the immediate and delayed feedback components were as follows:

$$V_{Ik+1}(s_k) = V_{Ik}(s_k) + \alpha^I(r_{Ik} - V_{Ik}(s_k)) \quad (1)$$

$$V_{Dk+1}(s_k) = V_{Dk}(s_k) + \alpha^D(r_{Dk} - V_{Dk}(s_k)) \quad (2)$$

The total value of a stimulus was the sum of the immediate and delayed values.

$$V_k(s_k) = V_{Ik}(sk) + V_{Dk}(sk) \quad (3)$$

Each value was initialized to 0 at $k = 0$.

For the baseline learning model, the learning rates for the immediate and delayed components were the same ($\alpha^I = \alpha^D = \alpha$).

For the differential weight model, the learning rates for the immediate and delayed components were the same but the estimated value of the delayed feedback was weighted less than the estimated value of the immediate feedback. The total value of a stimulus was the sum of the immediate and delayed values:

$$V_k(s_k) = V_{Ik}(sk) + \gamma V_{Dk}(sk) \quad (4)$$

where γ is the discount factor on the estimated value of the delayed component.

For the differential learning and weight model, the learning rates and choice weights for the immediate and delayed feedback were allowed to be different. This model was a combination of the differential learning model and the differential weight model.

The decision process itself was described by the DDM [16,79] with one boundary for correct responses and one boundary for incorrect responses. In each trial, the drift rate was defined as the difference between the values from the RL model. When the difference was high, the drift rate was higher, leading to more accurate and faster responses. On each trial, the drift rate was defined as:

$$v \sim d(V_{k,correct}(s_k) - V_{k,incorrect}(s_k)) \quad (5)$$

To compare models, we used the Widely Applicable Information Criterion (WAIC) [80]. This criterion is especially useful when the posterior distribution is not Normal. We computed WAIC for each subject and each model.

Model comparison provides only relative performance among models. To check whether the RL model captured the data well, we conducted a posterior predictive check. For each subject, we used the mean of the posterior distribution to generate 100 simulated choices and RTs using their actual sequence of trials. We computed choice accuracy at three levels: (1) at the trial-level when averaging across all subjects (Fig. 5A, B), (2) at the subject-level when averaging across all trials (Fig. S12), (3) overall level when averaging across both trials and subjects (Fig. S13) and response times across both trials and subjects (Fig. S14) [79]. We excluded trials where the two options were identical.

Priors

We chose weakly informative priors so as to have minimal influence on the posterior [81].

$$\alpha^I \sim B(1,1),\ 0 \leq \alpha^I \leq 1 \quad (6)$$

$$\alpha^D \sim B(1,1),\ 0 \leq \alpha^D \leq 1 \quad (7)$$

$$a \sim N(2,1),\ a \geq 0 \quad (8)$$

$$t \sim U(0,\ minRT) \quad (9)$$

$$d \sim N(0,\ 2),\ d \geq 0 \quad (10)$$

$$\gamma \sim B(1,1),\ 0 \leq \gamma \leq 1 \quad (11)$$

729

730

731 *Fitting procedure*

732 We used Stan to find the best fitting parameters for each subject separately, using the Monte-
733 Carlo Markov Chain sampling method. For each subject, four chains were run in parallel. Each
734 chain consisted of 4,000 samples out of which 2,000 were warm-up samples. We computed R-hat
735 of all parameters to assess model convergence. This method compares the between- and within-
736 chain estimates for model parameters. The maximum R-hat was 1.01 indicating model
737 convergence [81].

738

739 *Parameter recovery*

740 To test whether parameters of the models could be recovered well, we simulated one dataset for
741 176 subjects using the mean posterior of the fitted parameter values and the experimental trials
742 encountered by each subject in the task. We then fit the model on this simulated data. We found
743 large and significant correlations between the true and fitted parameter values and no significant
744 bias in the recovery of the model parameters (Fig. S12).

745

746 *Model recovery*

747 To test whether the differential learning model and the differential weight model could be
748 distinguished from one another, we examined model recovery. We simulated one dataset for 176
749 participants using the mean posteriors of the fitted parameter values and the experimental trials
750 encountered by each participant in the task for each model. We then fit each model on these
751 simulated datasets and compared the model fits using WAIC to determine whether the model

used to generate the data was also the best-fitting model. To check whether more trials leads to better model recovery, we also simulated datasets for each model using double the number of trials.

The model recovery analyses show that the differential weight model tends to mimic the differential learning model even when the differential learning model generated the data, but not the other way around (Simulated: Differential Learning Model, Fitted: 53% Differential Learning Model, 47% Differential Weight Model; Simulated: Differential Weight Model, Fitted: 72% Weight Model, 28% Differential Learning Model; Fig. S22). This is true even when using double the number of trials in the experiment (Simulated: Differential Learning Model, Fitted: 65% Differential Learning Model, 35% Differential Weight Model; Simulated: Differential Weight Model, Fitted: 71% Weight Model, 29% Differential Learning Model; Fig. S22).

Acknowledgments

We thank Adam Foa, Malia Young, Sumedha Goyal, and Walker Daniel for research assistance. National Science Foundation grant 2333979 (IK).

Preregistrations:

The preregistration for Study 2 is available at OSF: https://osf.io/mkgqy for Study 3 is available at OSF: https://osf.io/37qa2 and for Study 4 is available at OSF: https://osf.io/tuv48.

Author Contributions:

Conceptualization: DP

Methodology: MC, DP, IK

Investigation: MC, DP

Visualization: MC, DP

Funding acquisition: IK, DP

Project administration: IK, DP

Supervision: IK

Writing – original draft: MC, DP, IK

Writing – review & editing: MC, DP, IK

References

1. Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology. 2012 Dec 1;22(6):1075–81.

2. Glimcher PW. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. PNAS. 2011 Sept 13;108(Supplement 3):15647–54.

3. Daw ND, Doya K. The computational neurobiology of learning and reward. Current Opinion in Neurobiology. 2006 Apr;16(2):199–204.

4. Sutton RS, Barto AG. Reinforcement learning: An introduction, 2nd ed. Cambridge, MA, US: The MIT Press; 2018. xxii, 526 p. (Reinforcement learning: An introduction, 2nd ed).

5. Lee AS, Duman RS, Pittenger C. A double dissociation revealing bidirectional competition between striatum and hippocampus during learning. PNAS. 2008 Nov 4;105(44):17163–8.

6. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal Difference Models and Reward-Related Learning in the Human Brain. Neuron. 2003 Apr 24;38(2):329–37.

7. Schultz W. Subjective neuronal coding of reward: temporal value discounting and risk. European Journal of Neuroscience. 2010;31(12):2124–35.

8. Gershman SJ. The Successor Representation: Its Computational Logic and Neural Substrates. J Neurosci. 2018 Aug 15;38(33):7193–200.

9. Rubin JE, Vich C, Clapp M, Noneman K, Verstynen T. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. Eur J Neurosci. 2021 Apr;53(7):2234–53.

10. Witkowski PP, Park SA, Boorman ED. Neural mechanisms of credit assignment for inferred relationships in a structured world. Neuron. 2022 Aug 17;110(16):2680-2690.e9.

11. Zauberman G, Kim BK, Malkoc SA, Bettman JR. Discounting Time and Time Discounting: Subjective Time Perception and Intertemporal Preferences. Journal of Marketing Research. 2009 Aug 1;46(4):543–56.

12. Frederick S, Loewenstein G. Time Discounting and Time Preference: A Critical Review. Journal of Economic Literature. 2002;

13. Ericson KM, Laibson D. Intertemporal choice. In: Handbook of Behavioral Economics: Applications and Foundations 1 [Internet]. North-Holland; 2019 [cited 2025 Nov 25]. p. 1–67. Available from: https://www.sciencedirect.com/science/chapter/handbook/abs/pii/S2352239918300253

14. Kable JW. Valuation, Intertemporal Choice, and Self-Control. In: Glimcher PW, Fehr E, editors. Neuroeconomics. Academic Press; 2014. p. 173–92.

15. Dai J, Busemeyer JR. A probabilistic, dynamic, and attribute-wise model of intertemporal choice. Journal of Experimental Psychology: General. 2014;143(4):1489–514.

16. Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci. 2010 Oct;13(10):1292–8.

17. Shevlin BRK, Krajbich I. Attention as a source of variability in decision-making: Accounting for overall-value effects with diffusion models. Journal of Mathematical Psychology. 2021;105.

18. Cockburn J, Collins AGE, Frank MJ. A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. Neuron. 2014 Aug 6;83(3):551–7.

19. Murty VP, DuBrow S, Davachi L. The simple act of choosing influences declarative memory. J Neurosci. 2015 Apr 22;35(16):6255–64.

20. Chambon V, Théro H, Vidal M, Vandendriessche H, Haggard P, Palminteri S. Information about action outcomes differentially affects learning from self-determined versus imposed choices. Nat Hum Behav. 2020 Oct;4(10):1067–79.

21. Pupillo F, Bruckner R. Signed and unsigned effects of prediction error on memory: Is it a matter of choice? Neuroscience & Biobehavioral Reviews. 2023 Oct 1;153:105371.

22. Chen F, Zheng J, Wang L, Krajbich I. Attribute latencies causally shape intertemporal decisions. Nat Commun. 2024 Apr 5;15(1):2948.

23. Maier SU, Raja Beharelle A, Polanía R, Ruff CC, Hare TA. Dissociable mechanisms govern when and how strongly reward attributes affect decisions. Nat Hum Behav. 2020 Sept;4(9):949–63.

24. Sullivan NJ, Huettel SA. Healthful choices depend on the latency and rate of information accumulation. Nat Hum Behav. 2021 Dec;5(12):1698–706.

25. Weber EU, Johnson EJ, Milch KF, Chang H, Brodscholl JC, Goldstein DG. Asymmetric Discounting in Intertemporal Choice: A Query-Theory Account. Psychol Sci. 2007 June 1;18(6):516–23.

26. Zhang Z, Wang S, Good M, Hristova S, Kayser AS, Hsu M. Retrieval-constrained valuation: Toward prediction of open-ended decisions. Proceedings of the National Academy of Sciences. 2021 May 18;118(20):e2022685118.

27. Foerde K, Shohamy D. Feedback timing modulates brain systems for learning in humans. J Neurosci. 2011 Sept 14;31(37):13157–67.

28. Knowlton BJ, Mangels JA, Squire LR. A Neostriatal Habit Learning System in Humans. Science. 1996 Sept 6;273(5280):1399–402.

29. Foerde K, Race E, Verfaellie M, Shohamy D. A Role for the Medial Temporal Lobe in Feedback-Driven Learning: Evidence from Amnesia. J Neurosci. 2013 Mar 27;33(13):5698–704.

30. Pavlov IP. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford, England: Oxford Univ. Press; 1927. xv, 430 p. (Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex).

31. Bangasser DA, Waxler DE, Santollo J, Shors TJ. Trace conditioning and the hippocampus: the importance of contiguity. J Neurosci. 2006 Aug 23;26(34):8702–6.

32. Rescorla R, Wagner A. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. undefined [Internet]. 1972 [cited 2021 Apr 23]; Available from: /paper/A-theory-of-Pavlovian-conditioning-%3A-Variations-in-Rescorla/afaf65883ff75cc19926f61f181a687927789ad1

33. Ratcliff R, Frank MJ. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. Neural Comput. 2012 May;24(5):1186–229.

34. Konovalov A, Krajbich I. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. Nat Commun. 2016 Aug 11;7(1):12438.

35. Fontanesi L, Gluth S, Spektor MS, Rieskamp J. A reinforcement learning diffusion decision model for value-based decisions. Psychon Bull Rev. 2019 Aug 1;26(4):1099–121.

36. Pedersen ML, Frank MJ. Simultaneous Hierarchical Bayesian Parameter Estimation for Reinforcement Learning and Drift Diffusion Models: a Tutorial and Links to Neural Data. Comput Brain Behav. 2020 Dec;3(4):458–71.

37. Shevlin BRK, Smith SM, Hausfeld J, Krajbich I. High-value decisions are fast and accurate, inconsistent with diminishing value sensitivity. Proc Natl Acad Sci U S A. 2022 Feb 8;119(6):e2101508119.

38. Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MFS, Behrens TEJ. Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci. 2012 Mar;15(3):470–6.

39. Pirrone A, Azab H, Hayden BY, Stafford T, Marshall JAR. Evidence for the speed–value trade-off: Human and monkey decision making is magnitude sensitive. Decision. 2018;5(2):129–42.

40. Shenhav A, Karmarkar UR. Dissociable components of the reward circuit are involved in appraisal versus choice. Sci Rep. 2019 Feb 13;9(1):1958.

41. Yang X, Krajbich I. Webcam-based online eye-tracking for behavioral research. Judgment and Decision Making. 2021 Nov;16(6):1485–505.

42. Busemeyer JR, Townsend JT. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychol Rev. 1993 July;100(3):432–59.

43. Diederich A. Dynamic stochastic models for decision making under time constraints. Journal of Mathematical Psychology. 1997;41(3):260–74.

44. Roe RM, Busemeyer JR, Townsend JT. Multialternative decision field theory: a dynamic connectionist model of decision making. Psychol Rev. 2001 Apr;108(2):370–92.

45. Konovalov A, Krajbich I. Revealed strength of preference: Inference from response times. Judgment and decision making [Internet]. 2019 July [cited 2025 Nov 25];14(4). Available from: https://par.nsf.gov/biblio/10288586-revealed-strength-preference-inference-from-response-times

46. Ratcliff R, McKoon G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. Neural Comput. 2008 Apr;20(4):873–922.

47. Polanía R, Krajbich I, Grueschow M, Ruff CC. Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. Neuron. 2014 May 7;82(3):709–20.

48. Cavanagh JF, Wiecki TV, Kochar A, Frank MJ. Eye Tracking and Pupillometry are Indicators of Dissociable Latent Decision Processes. J Exp Psychol Gen. 2014 Aug;143(4):1476–88.

49. Papoutsaki A. Scalable Webcam Eye Tracking by Learning from User Interactions. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2025 Nov 25]. p. 219–22. (CHI EA '15). Available from: https://doi.org/10.1145/2702613.2702627

50. Semmelmann K, Weigelt S. Online webcam-based eye tracking in cognitive science: A first look. Behav Res Methods. 2018 Apr;50(2):451–65.

51. Herrnstein RJ, Prelec D. Melioration: A Theory of Distributed Choice. The Journal of Economic Perspectives. 1991;5(3):137–56.

52. Herrnstein RJ. Behavior, Reinforcement and Utility. Psychol Sci. 1990 July 1;1(4):217–24.

53. Herrnstein RJ. Experiments on stable sub-optimality in individual behavior. American Economic Review. 1991;360–4.

54. Heyman GM, Herrnstein RJ. More on Concurrent Interval-Ratio Schedules: A Replication and Review. Journal of the Experimental Analysis of Behavior. 1986;46(3):331–51.

55. Vaughan Jr. W. Melioration, Matching, and Maximization. Journal of the Experimental Analysis of Behavior. 1981;36(2):141–9.

950 56. Herrnstein RJ, Loewenstein GF, Prelec D, Vaughan Jr. W. Utility maximization and
951 melioration: Internalities in individual choice. Journal of Behavioral Decision Making.
952 1993;6(3):149–85.

953 57. Herrnstein, R.J. The Matching Law: Papers in Psychology and Economics. Rachlin H, D. I.
954 Laibson, editors. Harvard University Press; 1997.

955 58. Prelec D. Consuming at the Wrong Rate: Lessons from the Harvard Game. In: Sustainable
956 Consumption: Multi-disciplinary Perspectives In Honour of Professor Sir Partha Dasgupta |
957 Oxford Academic [Internet]. Oxford University Press; 2014 [cited 2025 Nov 25]. Available
958 from: https://academic.oup.com/book/32654/chapter-
959 abstract/270592248?redirectedFrom=fulltext

960 59. Gureckis TM, Love BC. Short-term gains, long-term pains: How cues about state aid
961 learning in dynamic environments. Cognition. 2009 Dec;113(3):293–313.

962 60. Neth H, Sims CR, Gray WD. Melioration Dominates Maximization: Stable Suboptimal
963 Performance Despite Global Feedback. :6.

964 61. Tunney RJ, Shanks DR. A re-examination of melioration and rational choice. J Behav Decis
965 Making. 2002 Oct;15(4):291–311.

966 62. Balasubramani PP, Diaz-Delgado J, Grennan G, Alim F, Zafar-Khan M, Maric V, et al.
967 Distinct neural activations correlate with maximization of reward magnitude versus
968 frequency. Cereb Cortex. 2023 May 9;33(10):6038–50.

969 63. Sims CR, Neth H, Jacobs RA, Gray WD. Melioration as rational choice: Sequential
970 decision making in uncertain environments. Psychological Review. 20121210;120(1):139.

971 64. Fiorillo CD, Newsome WT, Schultz W. The temporal precision of reward prediction in
972 dopamine neurons. Nature Neuroscience. 2008 Aug;11(8):966–73.

973 65. Kobayashi S, Schultz W. Influence of Reward Delays on Responses of Dopamine Neurons.
974 J Neurosci. 2008 July 30;28(31):7837–46.

975 66. Foerde K, Shohamy D. The role of the basal ganglia in learning and memory: insight from
976 Parkinson's disease. Neurobiol Learn Mem. 2011 Nov;96(4):624–36.

977 67. Lighthall NR, Pearson JM, Huettel SA, Cabeza R. Feedback-Based Learning in Aging:
978 Contributions and Trajectories of Change in Striatal and Hippocampal Systems. J Neurosci.
979 2018 Sept 26;38(39):8453–62.

980 68. Bakkour A, Palombo DJ, Zylberberg A, Kang YH, Reid A, Verfaellie M, et al. The
981 hippocampus supports deliberation during value-based decisions. Kahnt T, Frank MJ,
982 Fellows LK, Gluth S, editors. eLife. 2019 July 3;8:e46080.

983 69. Newell BR, Dunn JC, Kalish M. Systems of category learning: Fact or fantasy? In: The
984 psychology of learning and motivation: Advances in research and theory, Vol 54. San

Diego, CA, US: Elsevier Academic Press; 2011. p. 167–215. (The psychology of learning and motivation).

70. Gabaix X, Laibson D. Myopia and Discounting [Internet]. National Bureau of Economic Research; 2017 [cited 2025 Nov 25]. (Working Paper Series). Available from: https://www.nber.org/papers/w23254

71. Gregorios-Pippas L, Tobler PN, Schultz W. Short-Term Temporal Discounting of Reward Value in Human Ventral Striatum. Journal of Neurophysiology. 2009 Mar 1;101(3):1507–23.

72. Orquin JL, Mueller Loose S. Attention and choice: A review on eye movements in decision making. Acta Psychologica. 2013 Sept 1;144(1):190–206.

73. Chen DL, Schonger M, Wickens C. oTree—An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance. 2016 Mar 1;9:88–97.

74. de Leeuw JR. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. Behav Res Methods. 2015 Mar;47(1):1–12.

75. Kirchner WK. Age differences in short-term retention of rapidly changing information. J Exp Psychol. 1958 Apr;55(4):352–8.

76. Wixted JT. The forgotten history of signal detection theory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2020;46(2):201–33.

77. Hautus MJ. Corrections for extreme proportions and their biasing effects on estimated values of d′. Behavior Research Methods, Instruments & Computers. 1995;27(1):46–51.

78. Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behavior Research Methods, Instruments & Computers. 1999;31(1):137–49.

79. Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion Decision Model: Current Issues and History. Trends Cogn Sci. 2016 Apr;20(4):260–81.

80. Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. J Mach Learn Res. 2010 Dec 1;11:3571–94.

81. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, Third Edition. CRC Press; 2013. 677 p.

Supporting Information for

Delayed reward information is underweighted in reinforcement learning with dispersed feedback

Miruna Cotet[1,2*], David Poensgen[3*], Ian Krajbich[4+]

[1] The Ohio State University, Columbus, US
[2] Complexity Science Hub, Vienna, Austria
[3] Goethe University, Frankfurt, Germany
[4] University of California, Los Angeles, US

*These authors contributed equally.
+Corresponding author: Ian Krajbich

Email: krajbich@ucla.edu

This PDF file includes:

- Supporting text
- Figures S1 to S24
- Tables S1 to S19
- Materials

Supplementary Text

Supplementary Methods

Study 1: Colors

Subjects

The experiment was conducted at FLEX (Frankfurt Laboratory for Experimental Economic Research) in February and June 2018. 159 subjects (102 for the active learning condition and 57 for the passive learning condition) were recruited using ORSEE (*1*). All subjects signed a written consent, and the studies received approval from the Joint Ethics Committee for Economics and Business Administration of Goethe University Frankfurt and Gutenberg University Mainz.

Design

Subjects' task in the experiment was to learn the values of 6 abstract stimuli. The stimuli were 6 different colors. These values were initially unknown, and could be learned by sampling. The experiment consisted of 105 trials. In each trial, subjects were presented with two stimuli and could choose one. Each stimulus generated a specific number of points. These points were displayed after each round, allowing subjects to learn throughout the experiment. Payment depended on total points earned, giving subjects an incentive to learn the relative values as quickly and precisely as possible.

Feedback for each choice was not given at once, but split into two components: one shown directly after the choice, one shown with one trial delay. Directly after a choice, the immediate reward was displayed with a clear association with the stimulus chosen. One trial later, the subject would earn the delayed reward for this same choice, again clearly displayed in association with the stimulus (and alongside the immediate feedback of the current trial's choice).

A mixed design was used. Subjects were randomly assigned to one of two experimental conditions. Each of the two groups had to learn a different reward vector associated with the 6 different stimuli. The total values were evenly spaced from 8 to 18. For one group the following levels for the underlying rewards were used (immediate reward, delayed reward): (11, 7), (6, 10), (9, 5), (4, 8), (7, 3), (2, 6) while for the second group the following levels were used: (7, 11), (10, 6), (5, 9), (8, 4), (3, 7), (6, 2). These values were split so that, the immediate reward was 4 points higher or lower than the delayed reward. This arrangement maximized the number of rounds in which subjects had to choose between options close in value, but with opposite temporal profiles.

A small random reward from a uniform distribution consisting of either 1, 2, 3 or 4 points was added to each underlying reward to make learning more difficult. Subjects were informed about this in the instructions. However, the variance was small relative to the value differences between the colors. For options 2 points apart, the chance of the worse option appearing better was only .14 after a single draw for each, and quickly shrank with additional sampling. For stimuli 4 points apart, this probability was .02, and reversal was impossible for larger value differences.

The sequence of 105 decisions was split Into 5 blocks, each with 21 decisions corresponding to all possible two-stimulus pairings. In 6 rounds both options were the same. These choice sets were included for several reasons. Most importantly, they forced subjects to sample each color in regular intervals. This greatly limited the need for deliberate exploration and ensure that subjects would have to sample each color at least once per block. Consequently, even when choices were heavily biased, they would still continuously generate ample evidence to correct this bias.

The trade-off between exploration and exploitation was very limited in this task. This was partly because variance was so small, partly because choice sets rotated in a way which essentially removed any need for deliberate exploration. It ensured that inferior options would soon be sampled anyway: either when up against an even worse option, or itself. The causal structure was explicitly revealed to subjects, and values depended only on the stimuli, so that subjects did not need to infer or memorize states of the world or any rules. Subjects clearly saw which choice had which consequences. Credit attribution was therefore not an issue here.

Block randomization was used to assign subjects to one of the two groups. A pre-fixed sequence of placeholders for stimulus presentation was used which ensured that each two-stimulus combination appeared five times. The pre-fixed sequence was chosen such that if a stimulus was available in round t, it was not available in rounds t + 1 or t + 2. This ensured that subjects had seen as many realizations of the delayed reward as of the immediate reward whenever a stimulus was available. The rule further guaranteed that while the stimulus was available, neither the immediate nor the delayed reward could be on screen. Within each subject, the following aspects of the sequence were randomized: which

stimulus took which place in the sequence, which average reward vector corresponded to which stimulus, left and right positions of the stimuli per round, and which of two sets of stimuli was used.

To address the question of whether passive as opposed to active learning would reduce the bias, an additional condition of the experiment allowed subjects to learn from others' feedback for the first part of the experiment before making their own choices. The passive treatment kept all rules and mechanics of the main experiment with one change: in the first 63 rounds, subjects did not make any decisions, but could learn passively from feedback shown on their screen. The feedback shown was based on the decisions of a matched partner from the main experiment. The feedback and timing of the feedback was the same as that of the matched partner, except that the subject could not see the foregone choice option on the choice screen. This was both to limit similarity to a choice situation, and to not convey information about the beliefs of the subject in the active learning condition to the subject in the passive learning condition. Subjects were not explicitly informed how the data had been generated. Both subjects entered the final 42 rounds with exactly the same information. The sequence of choice sets in these rounds was again kept identical. Instructions were unchanged, with an added section explaining the learning phase.

Procedure

The experiment was programmed in oTree (*78*). An important feature of the experiment was complete transparency of its causal structure. Only the reward vectors were explicitly not disclosed to subjects; they received no information on their range, average or similar. All other mechanics of the experiment were clearly explained in the instructions. Before starting the main task, subjects played an interactive tutorial, which replicated the main task with slight modifications: Detailed on-screen explanations of all rules could be shown and hidden at any time. Instead of 6 colors, 4 shades of gray were used; after a few rounds had been played, their respective values (immediate, delayed) were revealed to make completely transparent how points are generated. The noise terms added to the rewards could be turned on and off at will during the tutorial.

The subjects then completed the main task. Once new options had appeared, subjects had a time limit of 10 seconds per decision, represented by a shrinking bar. If time ran out, a choice was made at random, and a penalty of 5 points deducted. After completing the main task, subjects were asked which of the colors gave the most points, the second most points, and the least points (both components combined). The same color could not be named twice; the questions were not incentivized.

At the end of the task, subjects completed an additional memory questionnaire. The first survey asked: 'Which of these images was associated with the highest (second, third, fourth, fifth highest or lowest) number of points?'. This question was asked only for the best, second best and worst option. The second survey consisted of a Likert scale asking: 'For which image was the first reward larger (smaller or equal) than the second reward'. This question was asked only for the second, third and fourth best options. The third survey asked subjects to estimate the number of points for each image. This question was asked only for the best and second-best options.

Afterwards, the subjects also completed a task to elicit their intertemporal preferences. This was done using a staircase estimator; this procedure is validated against incentivized decisions (*2*). In three series of questions, subjects were asked how they would decide between 100 euros today or x in one month; 100 euros today or x in six months; 100 euros in one month or x in six months. By increasing or decreasing x from question to question, effectively performing bisection search, indifference points in the range 100–132 euros were obtained for each time horizon. By nature of the staircase procedure, more extreme values are censored.

After the experiment, subjects were paid 0.05 euros for each point scored above 1,800 in the task, and nothing if below that threshold. This created steep incentives in the relevant region: Each of the 75 non-degenerate decisions effectively had a stake between 0.10 euros and 0.50 euros. Subjects could score between 1,715 – 2,065 points (ignoring penalties for timeouts), and random behavior was expected to earn 1,890. Subjects were paid an additional 2 euros for completion of the remaining questionnaire. Sessions lasted around 50 minutes. Subjects earned 10.78 euros on average.

Study 2: Patterns Task

Subjects

A total of 374 subjects were recruited from The Ohio State University between February and July 2021. All subjects gave written consent, and the study was approved by the OSU Institutional Review Board (2013B0583). Subjects could

participate in the experiment only if they passed a short initial eye-tracking calibration, a control questionnaire and a second longer eye-tracking calibration. 178 (47%) subjects failed the initial calibration and were paid $2, 49 (13%) subjects failed the control questionnaire and were paid $3 and 23 (6%) subjects failed the second calibration and were paid $4.

A total of 124 subjects completed the experiment. Subjects were paid based on their choices in the task. The minimum payment was $9 for completing the task and the maximum payment was $14. Subjects earned an average of $10. Subjects were paid 4 cents for each point earned above the 1,505 points threshold and could earn a maximum of 350 points. Subjects also lost 5 points if they did not make a decision within 3 seconds.

Our target sample size was 80 subjects who satisfy our exclusion restrictions, 40 for each group. We determined the sample size based on a previous pilot study. Our goal was to obtain at least .80 power to detect a significant Likelihood Ratio Test for whether the immediate reward has a higher coefficient than the delayed reward in a mixed effects logistic regression in which the outcome variable is the choice of the left stimulus and the predictors were the difference between the left and right option in the experienced average immediate reward and delayed reward.

Design

The design was similar to the colors task, with a few modifications that allowed for better eye-tracking measures. First of all, the stimuli were two sets of 6 abstract black and white art images. We choose these images to be similar in salience. The salience was computed using the Graph-based Visual Saliency toolbox in MATLAB (*3*). Secondly, we separated the choice and feedback stages into two separate screens. This was done in order to make it easier to measure subjects' attention to the feedback information. A randomly generated time interval between two and six seconds was added after each choice and feedback screen. Thirdly, subjects had a maximum of 3 seconds to make their choice, instead of 10 seconds. If no choice was made within this time, one stimulus was chosen randomly by the computer. The feedback screen was presented for only 2 seconds. Due to issues with timing on web browsers, the feedback was on average presented for 2.5 seconds. Lastly, the small random reward from a uniform distribution consisting of either 0, 1, 2 or 3 points was added to each underlying reward instead of either 1, 2, 3 or 4 points.

Gaze position was measured using subjects' webcams. We recorded the horizontal (x) and vertical (y) gaze position at each moment in time during the choice and feedback phases of each trial. We defined the Area of Interest (AOI) for the immediate reward as the top half of the screen and the AOI for the delayed reward as the bottom half of the screen.

Procedure

Potential subjects received a link which they could access to complete the experiment. First, subjects completed a consent form. Then they answered a few demographic questions and questions about preferred online payment method. Second, subjects completed a short eye-tracking calibration. Third, they read the instructions and completed a control questionnaire to check their understanding of the instructions. To advance to the second calibration, they had to answer 4 out of 5 questions correctly. After completing the second calibration, they advanced to the experiment.

A short validation phase was given after 21, 42, and 84 trials. After 63 trials subjects completed another calibration and validation. This was done to ensure high eye-tracking data quality. If subjects failed all 3 short validations, they were excluded. Each of the validation checks consisted of showing 3 dots at different positions on the screen that the subjects had to fixate. A subject failed the validation if none of the 3 dots were considered fixated within a certain radius around the dot.

At the end of the choice phase, subjects completed 3 short memory surveys. The first survey asked: 'Which of these images was associated with the highest (second, third, fourth, fifth highest or lowest) number of points?'. The second survey consisted of a Likert scale asking: 'For which image was the first payoff larger (smaller or equal) than the second payoff'. The third survey asked subjects to estimate the number of points for each image.

At the end of the experiment, subjects were shown their final reward and were paid with their preferred online method. Experiments lasted 37 minutes on average.

The analyses were preregistered at OSF (osf.io/mkgqy).

Study 3: Patterns Task Position Feedback Reversed

Subjects

A total of 169 subjects were recruited from Prolific between February and March 2024. All subjects gave their consent, and the studies received approval from the OSU Institutional Review Board (2023E1158). Subjects could participate in the experiment only if they passed a control questionnaire. 112 (66%) subjects failed the control questionnaire. A total of 57 subjects completed the experiment. Subjects were paid based on their choices in the task. The minimum payment was $9 for completing the task and the maximum payment was $14. Subjects earned an average of $10. Subjects were paid 4 cents for each point earned above the 1,505 points threshold and could earn a maximum of 350 points. Subjects also lost 5 points if they did not make a decision within 3 seconds.

Our target sample size was 40 subjects who satisfy our exclusion restrictions. We determined the sample size based on Study's 2 data. Our goal was to obtain at least .80 power to detect a significant Likelihood Ratio Test for whether the immediate reward has a higher coefficient than the delayed reward in a mixed effects logistic regression in which the outcome variable is the choice of the left stimulus and the predictors were the difference between the left and right option in the experienced average immediate and delayed rewards.

Design

The design was the same as the patterns task, except for the position of the feedback on the screen. In Study 1, the immediate reward was always shown on the top half of the screen while the delayed reward was always shown on the bottom half of the screen. Here, we reversed these positions such that the immediate reward was always on the bottom half of the screen while the delayed reward was always on the top half of the screen. This allowed us to test whether the immediacy bias was instead just a spatial bias.

Procedure

The procedure was the same as for Study 1 except that subjects on Prolific did not have to answer any demographic questions or select a preferred payment method. We also did not exclude subjects if they failed the eye-tracking calibration.

The following analyses were preregistered at OSF (osf.io/37qa2).

Study 4: In-lab Eye-tracking Study

Subjects

The experiment was conducted at UCLA (University of California Los Angeles Behavioral Research Lab) in May and November 2024. 57 subjects were recruited for the study. All subjects gave their consent, and the studies received approval from the University of California Los Angeles Review Board. 10 subjects were excluded because their accuracy on the congruent choice sets was below 60%.

Design

The design was the same as the online patterns task (Study 2), except we varied the position of the feedback on the screen between participants. For half of the participant, the immediate reward was always shown on the top half of the screen while the delayed reward was always shown on the bottom half of the screen, while for the other half we reversed these positions such that the immediate reward was always on the bottom half of the screen while the delayed reward was always on the top half of the screen. This allowed us to test whether the immediacy bias was instead just a spatial bias.

Procedure

The procedure was the same as the online patterns task (Study 2) except that subjects did the task in the lab. Subjects also did additional tasks as in Study 1. We replaced the n-back working memory task with a visual memory task, namely the change localization task (*4*). We also made a change to the declarative memory task. For the point estimation task, subjects had to estimate separately the average points for the immediate and delayed reward.

Subjects were eye-tracked while performing the task. Monocular eye tracking data were collected with a remote EyeLink 1000 Plus system (SR Research Ltd., Mississauga, Ontario, Canada), with a sampling frequency of 500 or

1000 Hz. Before the start of each trial, subjects had to fixate a central fixation cross to ensure that they began each trial fixating on the same location. We recorded the horizontal and vertical gaze position at each moment in time during the trial. The AOIs corresponded to the top and bottom part of the screen for the feedback and the left and right part of the screen for the choice. Similar to Study 2 and 3, subjects had a maximum of 3 seconds to make their choice. If no choice was made within this time, one stimulus was chosen randomly by the computer. The feedback screen was presented for only 2 seconds.

The following analyses were preregistered at OSF (osf.io/tuv48).

Supplementary Results

Colors and Patterns Tasks

Response Times

The learning bias in favor of options with higher immediate rewards was also evident in RT-inferred indifference points. If subjects overweigh immediate rewards, they should be slowest to decide when the descending option is slightly worse than the ascending option. This would correspond to the subjects' average indifference point. To test this hypothesis, we used a mixed-effects quadratic regression with logarithm of RT as the outcome variable. As expected, we found a negative quadratic coefficient for difference in total reward between descending and ascending options (Study 1: $\beta = -0.002, 95\%CI = [-0.003, -0.002], p < 10^{-15}$; Study 2: $\beta = -0.001[-0.002, -0.001], p < 10^{-11}$; Fig. S5, Table S5). Moreover, the estimated peak of this RT curve was at -2.32 for the colors task and -0.81 for the patterns task, indicating that indeed subjects' average indifference point occurred when the descending option was slightly worse than the ascending option.

Passive versus Active Learning

For the passive learning conditions the sample size consisted of 57 subjects. Although there was only marginal evidence that passive learning reduced the immediacy bias when considering experienced average rewards, there was significant evidence for this reduction when considering choices of descending versus ascending options and error rates for congruent and incongruent choice sets.

There was only marginally significant evidence that the bias in the passive learning condition was lower than in the active learning condition when considering the experienced average immediate and delayed rewards. We used a regression of Choose Left on differences in the average immediate and delayed between the left and right options as well as a dummy for the passive learning condition and interactions of this dummy with the average immediate and delayed rewards. The coefficient for the interaction of the experienced average immediate reward and passive learning condition was marginally negative ($\beta_{Immediate:Passive} = -0.49, 95\%CI = [-1.04, 0.07], p = .085$; Table S2). The coefficient for the interaction of the experienced average delayed reward and passive learning condition was positive, but not significant ($\beta_{Delayed:Passive} = 0.15, 95\%CI = [-0.31, 0.62], p = .517$; Table S2).

We also checked for the immediacy bias by using regression of Choose Left on whether the options were ascending or descending, controlling for total rewards difference between left and right options. As before, we included a dummy for the passive learning condition and interactions of this dummy with whether the options were ascending or descending. There was evidence that the immediacy bias was lower in the passive learning condition when considering choice of descending versus ascending options. Choosing the right option when it was descending was less likely in the passive learning condition compared to the active learning condition as evidenced by a positive interaction coefficient between passive learning condition and whether the right stimulus was descending as opposed to ascending ($\beta = 0.95, 95\%CI = [0.23, 1.66], p = .009$; Table S1). The same was not true for when the left option was descending, though the direction of the effect was consistent with a lower bias ($\beta = -0.43, 95\%CI = [-1.12, 0.26], p = .218$; Table S1).

Incongruent choice sets had higher error rates in the active learning condition than the passive learning condition when looking at matched pairs of subjects (Active: $M = 0.53, SD = 0.30$; Passive: $M = 0.39, SD = 0.31, t(48) = -2.21, 95\%CI = [-0.28, -0.01], p = .031$). The opposite was true for error rates for congruent choice sets (Active: $M = 0.07, SD = 0.13$; Passive: $M = 0.16, SD = 0.24, t(48) = 2.02, 95\%CI = [0.00, 0.16], p = .049$) indicating

that the immediacy bias was lower in the passive learning condition compared with the active learning condition. For these analyses we used only the subjects that passed the exclusion criteria from Study 1. We excluded 15 subjects whose accuracy on trials with both options ascending or descending was below 60 percent.

Reversed Feedback Position

For the following analyses, we combined part of the data from Study 2 with the data from Study 3. We selected the first 40 odd-numbered participants from Study 2 (22 in Group 1, 18 in Group 2) and the first 40 participants from Study 3 (15 in Group 1, 25 in Group 2). We used the same exclusion criteria as before – 71 trials were excluded from Study 3 because subjects were too slow. This plan was pre-registered.

Behavior

For the same total reward level, the option with the higher immediate reward was more likely to be chosen compared to the option with the lower immediate reward (Fig. S16A). We confirmed these results using regressions of Choose Left on differences in the average immediate, delayed, and total rewards between the left and right options, as well as whether the options were ascending or descending. We used mixed-effects regressions with random intercepts and slopes at the subject level. For each trial we calculated the relevant average rewards seen by the subject up to that point in the experiment. We performed these regressions both when pooling the data from the two conditions and when using a dummy variable for the position condition and its interactions with the other variables in the model.

Choosing the left option was more likely when it was descending as opposed to ascending (Studies 2 and 3: $\beta_{LeftFalling} = 0.29, 95\%CI = [0.09, 0.50], p = .005$; Fig. S16C, Table S14) or when the right option was ascending rather than descending (Studies 2 and 3: $\beta_{RightFalling} = -0.39, 95\%CI = [-0.60, -0.19], p < 10^{-4}$; Fig. S16C, Table S14), controlling for the difference in total rewards. There was no main effect of position of the feedback on the choice of the left option (Studies 2 and 3: $\beta_{ImmediateBottom} = -0.007, 95\%CI = [-0.30, 0.29], p = .961$; Table S14). The bias in favor of descending options was even stronger when the immediate reward feedback was at the bottom of the screen compared to when it was at the top of the screen ($\beta_{LeftFalling:ImmediateBottom} = 0.32, 95\%CI = [0.03, 0.61], p = .033$; Fig. S16B, Table S14).

Moreover, congruent choice sets had lower error rates than incongruent choice sets (Studies 2 and 3: $\beta_{SetCongruent} = -0.14, 95\%CI = [-0.23, -0.05], p = .003$; Table S15). The position of the feedback on the screen did not make a difference on error rates (Studies 2 and 3: $\beta_{ImmediateBottom} = 0.01, 95\%CI = [-0.08, 0.10], p = .748$, $\beta_{SetCongruent:ImmediateBottom} = -0.07, 95\%CI = [-0.19, 0.06], p = .299$; Table S15).

Subjects put larger weights on immediate rewards than delayed rewards. When regressing choice on the immediate and delayed reward differences, the weight on the immediate reward was higher than the weight on delayed reward (Studies 2 and 3: $\beta_{Immediate} = 1.48, 95\%CI = [1.26, 1.69], p < 10^{-15}, \beta_{Delayed} = 1.01[0.80, 1.22], p < 10^{-15}$; Fig. S16B, Table S16). A Likelihood Ratio Test comparing the immediate and delayed coefficients was significant (Studies 2 and 3: $\chi^2(4, N = 80) = 291.53, p < 10^{-15}$). The position of the feedback did not make a difference (Studies 2 and 3: $\beta_{ImmediateBottom} = 0.08, 95\%CI = [-0.11, 0.28], p = .405, \beta_{Immediate:ImmediateBottom} = 0.004, 95\%CI = [-0.42, 0.43], p = .984, \beta_{Delayed:ImmediateBottom} = -0.04, 95\%CI = [-0.46, 0.37], p = .845$; Fig. S16B, Table S16).

To test whether this behavioral bias increased or decreased over the course of the experiments, we added interaction effects with trial number to the previous regression. The interaction of trial number and immediate reward was positive and significant (Studies 2 and 3: $\beta_{Immediate:Trial} = 0.34, 95\%CI = [0.26, 0.43], p < 10^{-14}$; Table S16), while the coefficient for the interaction of trial number and experienced delayed reward was also positive, but not significant (Studies 2 and 3: $\beta_{Delayed:Trial} = 0.06, 95\%CI = [-0.01, 0.14], p = .112$; Table S16). A Likelihood Ratio Test comparing the immediate and delayed interaction coefficients was significant (Studies 2 and 3: $\chi^2(4, N = 80) = 242.52, p < 10^{-16}$). This indicates that the immediacy bias increases over the course of the experiment. The position of the feedback did not make a difference (Studies 2 and 3: $\beta_{Immediate:Trial:ImmediateBottom} = -0.11, 95\%CI = [-0.28, 0.06], p = .205, \beta_{Delayed:Trial:ImmediateBottom} = -0.03, 95\%CI = [-0.19, 0.12], p = .662$; Table S16).

Response Times

Larger total |VD| and larger total OV decreased RT (Studies 2 and 3: $\beta_{|VD|} = -0.03[-0.04, -0.02], p < 10^{-5}, \beta_{OV} = -0.04[-0.06, -0.03], p < 10^{-7}$; Fig. S18, Table S17). Position of the feedback did not

make a difference ($\beta_{|VD|:ImmediateBottom} = 0.004[-0.02, 0.03], p = .723, \beta OV : ImmediateBottom = 0.0005[-0.03, 0.03], p = .970$; Fig. S18, Table S17). When we separated value into immediate and delayed rewards, we also found that both immediate and delayed |VD| decreased RT (Studies 2 and 3: $\beta_{VDImmediate} = -0.03[-0.04, -0.02], p < 10^{-7}; \beta_{VDDelayed} = -0.02[-0.03, -0.01], p = .001$; Figure S18, Table S18). The difference between immediate and delayed was significant (Studies 2 and 3: $\chi^2(1, N = 80) = 3.851, p = .05$). Position of the feedback did not make a difference ($\beta_{VDImmediate:ImmediateBottom} = -0.002[-0.02, 0.02], p = .838; \beta_{VDDelayed:ImmediateBottom} = -0.001[-0.02, 0.02], p = .918$; Fig. S18, Table S18). We did not find that immediate OV had a larger effect on RT than delayed OV (Studies 2 and 3: $\beta_{OVImmediate} = -0.03[-0.04, -0.02], p < 10^{-10}; \beta_{OVDelayed} = -0.03[-0.04, -0.02], p < 10^{-8}; \chi^2(1, N = 80) = 0.4542, p = .50$; Fig. S18, Table S18). Position of the feedback did not make a difference ($\beta_{OVImmediate:ImmediateBottom} = -0.01[-0.03, 0.01], p = .315; \beta_{OVDelayed:ImmediateBottom} = -0.008[-0.01, 0.03], p = .423$; Fig. S18, Table S18).

The learning bias in favor of options with higher immediate rewards was also evident in subjects' RT-inferred indifference points. We found a negative quadratic coefficient for difference in total reward between descending and ascending options (Studies 2 and 3: $\beta_{Total} = -0.001, 95\%CI = [-0.001, -0.001], p < 10^{-4}$; Fig. S5, Table S19). Moreover, the estimated peak of this RT curve was at -1.26 indicating that the subject's indifference point occurred when the descending option was slightly worse than the ascending option, as expected. The position of the feedback did not make a difference (Studies 2 and 3: $\beta_{Total:ImmediateBottom} = -0.001, 95\%CI = [-0.008, 0.006], p = .707, \beta_{TotalSq:ImmediateBottom} = -0.00006, 95\%CI = [-0.001, 0.001], p = .902$; Table S19).

Model

For the differential learning model, the learning rate for the immediate reward was higher than for the delayed reward for reversed feedback position condition (Study 3: $M_{Immediate} = 0.28, M_{Delayed} = 0.20, t(39) = 2.96, 95\%CI = [0.030.15], p = .005$; Fig. S16D). According to subject level WAIC, the model with differential learning provides a better fit for 57% of our subjects while the model with the same learning rate for both types of rewards fit 42% of our subjects better.

For the weight model, we found lower weight on the delayed rewards compared with immediate rewards (Study 3: $M_{Delayed} = 0.52, t(39) = -14.23, 95\%CI = [0.45, 0.59], p < 10^{-15}$).

Using all models (Baseline Model, Differential Learning Model, Differential Weight Model and Differential Learning and Weight Model), the model with same learning rates and equal decision weights provides a better fit for 35% subjects, the model with differential learning rate and equal weights provides a better fit for 15% subjects, the model with equal learning rates and different weights provides a better fit for 27% of subjects, while the model with both different learning rates and different weights provides a better fit for 23% subjects.

Gaze Bias

For these analyses, we used 18 subjects from Study 3, after excluding 22 subjects who did not pass the initial calibration or did not pass at least three out of four validation checks throughout the experiment. We did not exclude any trials. On the whole, subjects did not have a tendency to look more at the immediate reward compared to the delayed reward (Fig. S20C). In a mixed-effects regression of relative dwell proportion on immediate vs. delayed, the effect (i.e., intercept) was not significant when controlling for the size and type (ascending vs. descending) of reward ($\beta = -0.02, 95\%CI = [-0.13, 0.09], p = .733$; Table S6), and also not significant when controlling for the predicted values and prediction errors ($\beta = -0.03, 95\%CI = [-0.17, 0.11], p = .664$; Table S7).

Subjects were not more likely to fixate first to the immediate reward compared to the delayed reward. In a mixed-effects logistic regression of first fixation location (immediate vs. delayed) the effect (i.e., intercept) was not significant when controlling for the size and type of reward ($\beta = 0.31, 95\%CI = [-0.13, 0.75], p = .171$; Table S6) nor when controlling for the predicted values and prediction errors ($\beta = 0.02, 95\%CI = [-0.59, 0.63], p = .958$; Table S7).

The learning bias was significantly correlated with the first fixation bias but not with the dwell proportion bias. However, this was true only when using the behavioral bias calculated from the regressions (Dwell Proportion: $\rho(16) = .24, p = .33$; First Fixation: $\rho(16) = .49, p = .04$) but not when calculated from the RL model (Dwell Proportion: $\rho(16) = .11, p = .65$; First Fixation: $\rho(16) = .33, p = .19$; Fig. S20). However, in this study the correlations were in the expected direction.

In-lab Eye-tracking Study

We used the same exclusion criteria as before, results in 99 trials being excluded from Study 4 because subjects were too slow. This plan was pre-registered.

Behavior

For the same total reward level, the option with the higher immediate reward was more likely to be chosen compared to the option with the lower immediate reward (Fig. S17A). We confirmed these results using regressions of Choose Left on differences in the average immediate, delayed, and total rewards between the left and right options, as well as whether the options were ascending or descending. We used mixed-effects regressions with random intercepts and slopes at the subject level. For each trial we calculated the relevant average rewards seen by the subject up to that point in the experiment. We performed these regressions both when pooling the data from the two conditions and when using a dummy variable for the position condition and its interactions with the other variables in the model.

Choosing the left option was more likely when it was descending as opposed to ascending ($\beta_{LeftFalling} = 0.37, 95\%CI = [-0.07, 0.81], p = .096$; Fig. S17C, Table S14) or when the right option was ascending rather than descending ($\beta_{RightFalling} = -0.57, 95\%CI = [-1.03, -0.11], p = .016$; Fig. S17C, Table S14), controlling for the difference in total rewards. However, when including the difference in total experienced rewards, the effects reversed ($\beta_{LeftFalling} = -1.80, 95\%CI = [-2.48, -1.13], p < 10^{-6}$, $\beta_{RightFalling} = 1.74, 95\%CI = [1.07, 2.42], p < 10^{-6}$, $\beta_{ImmediateBottom} = -0.21, 95\%CI = [-0.68, 0.27], p = .393$ $\beta_{LeftFalling:ImmediateBottom} = 0.57, 95\%CI = [-0.42, 1.55], p = .258$, $\beta_{RightFalling:ImmediateBottom} = -0.52, 95\%CI = [-1.50, 0.46], p = .299$; Table S14). While this goes against the results of Studies 1-3, we also test for the bias in a different way below, using the points of the immediate and delayed rewards instead of whether an option was rising or falling, and find strong evidence for the immediacy bias. Position of the feedback did not have a significant effect.

Congruent choice sets had lower error rates than incongruent choice sets which is consistent with results from Studies 1-3 and indicate that a behavioral bias in favor of the immediate reward is present ($\beta_{SetCongruent} = -0.19, 95\%CI = [-0.31, -0.07], p = .002$; Table S15). The position of the feedback on the screen did not make a difference on error rates ($\beta_{ImmediateBottom} = 0.10, 95\%CI = [-0.02, 0.22], p = .110$, $\beta_{SetCongruent:ImmediateBottom} = -0.13, 95\%CI = [-0.30, 0.04], p = .144$; Table S15).

Subjects put larger weights on immediate rewards than delayed rewards. When regressing choice on the immediate and delayed reward differences, the weight on the immediate reward was higher than the weight on delayed reward ($\beta_{Immediate} = 1.49, 95\%CI = [1.20, 1.78], p < 10^{-15}, \beta_{Delayed} = -0.35[-0.55, -0.16], p < 10^{-3}$; Fig. S17B, Table S16). A Likelihood Ratio Test comparing the immediate and delayed coefficients was significant ($\chi^2(4, N = 47) = 578.17, p < 10^{-15}$). The position of the feedback did not make a difference ($\beta_{ImmediateBottom} = -0.22, 95\%CI = [-0.55, 0.10], p = .183, \beta_{Immediate:ImmediateBottom} = 0.15, 95\%CI = [-0.42, 0.72], p = .615, \beta_{Delayed:ImmediateBottom} = -0.11, 95\%CI = [-0.50, 0.28], p = .580$; Fig. S17B, Table S16).

To test whether this behavioral bias increased or decreased over the course of the experiments, we added interaction effects with trial number to the previous regression. The interaction of trial number and immediate reward was positive and significant ($\beta_{Immediate:Trial} = 0.29, 95\%CI = [0.18, 0.41], p < 10^{-6}$; Table S16), while the coefficient for the interaction of trial number and experienced delayed reward was negative and significant ($\beta_{Delayed:Trial} = -0.25, 95\%CI = [-0.35, -0.14], p < 10^{-5}$; Table S16). A Likelihood Ratio Test comparing the immediate and delayed interaction coefficients was significant ($\chi^2(4, N = 47) = 2922.2, p < 10^{-15}$). This indicates that the immediacy bias increases over the course of the experiment. The position of the feedback did not make a difference ($\beta_{Immediate:Trial:ImmediateBottom} = -0.002, 95\%CI = [-0.23, 0.23], p = .983, \beta_{Delayed:Trial:ImmediateBottom} = -0.04, 95\%CI = [-0.25, 0.17], p = .704$; Tables S16).

Response Times

Larger total |VD| and larger total OV decreased RT, but larger total |VD| was only marginally significant ($\beta_{|VD|} = -0.02[-0.040, 0.001], p = .067, \beta_{OV} = -0.03[-0.04, -0.01], p = .005$; Fig. S19, Table S17). Position of the immediate feedback at the bottom of the screen did not make a difference for |VD|, but it decreased RT for OV consistent with a stronger effect ($\beta_{|VD|:ImmediateBottom} = -0.02[-0.06, 0.02], p = .305, \beta_{OV:ImmediateBottom} = -0.04[-0.074, -0.003], p = .033$; Fig. S19, Table S17). When we separated

value into immediate and delayed rewards, we found that only immediate |VD| decreased RT, while delayed |VD| increases RT suggesting a strong bias towards considering the immediate rewards more than the delayed rewards ($\beta_{|VD|Immediate} = -0.04[-0.05, -0.02], p < 10^{-8}; \beta_{|VD|Delayed} = 0.01[0.002, 0.025], p = .022$; Figure S19, Table S18). The difference between immediate and delayed was significant ($\chi^2(1, N = 47) = 34.502, p < 10^{-8}$). Position of the immediate feedback at the bottom of the screen decreased RT further, but only for the immediate |VD|, while it did not make a difference for the delayed |VD| ($\beta_{|VD|Immediate:ImmediateBottom} = -0.05[-0.07, -0.03], p < 10^{-4}; \beta_{|VD|Delayed:ImmediateBottom} = -0.01[-0.01, 0.04], p = .244$; Fig. S19, Table S18). We also found that only immediate OV decreased RT, while delayed OV increased RT suggesting a strong bias towards considering the immediate rewards more than the delayed rewards. The difference between immediate and delayed was significant ($\beta_{OVImmediate} = -0.03[-0.04, -0.02], p < 10^{-7}; \beta_{OVDelayed} = 0.02[0.01, 0.03], p = .002; \chi^2(1, N = 47) = 32.693, p < 10^{-7}$; Fig. S19, Table S18). Position of the immediate feedback at the bottom of the screen decreased RT further for higher immediate OV and increased RT for higher delayed OV ($\beta_{OVImmediate:ImmediateBottom} = -0.04[-0.07, -0.02], p < 10^{-3}; \beta_{OVDelayed:ImmediateBottom} = 0.03[0.002, 0.048], p = .034$; Fig. S19, Table S18).

The learning bias in favor of options with higher immediate rewards was also evident in subjects' RT-inferred indifference points. We found a negative quadratic coefficient for difference in total reward between descending and ascending options ($\beta_{Total} = -0.007, 95\%CI = [-0.013, -0.002], p = .012$; Fig. S5D, Table S19). Moreover, the estimated peak of this RT curve was at -1.26 indicating that the subject's indifference point occurred when the descending option was slightly worse than the ascending option, as expected. The position of the immediate feedback at the bottom of the screen had an effect only on the linear coefficient shifting the indifference point to -5.25 increasing the observed bias ($\beta_{Total:ImmediateBottom} = -0.01, 95\%CI = [-0.025, -0.004], p = .007, \beta_{TotalSq:ImmediateBottom} = -0.0001, 95\%CI = [-0.001, 0.001], p = .771$; Table S19).

Model

For the differential learning model, the learning rate for the immediate reward was higher than for the delayed reward ($M_{Immediate} = 0.24, M_{Delayed} = 0.14, t(46) = 5.26, 95\%CI = [0.07, 0.15], p < 10^{-5}$; Fig. S17D). According to subject level WAIC, the model with differential learning provides a better fit for 60% of our subjects while the model with the same learning rate for both types of rewards fit 40% of our subjects better. When the immediate feedback was at the top (bottom), 36% (45%) of subjects were better fit by the model with differential learning compared to 64% (55%) by the model with the same learning rate. Thus, according to the model comparison, the learning bias in favor of the immediate reward was stronger when the immediate reward feedback was at the bottom of the screen.

For the weight model, we found lower weight on the delayed rewards compared with immediate rewards (Study 4: $M_{Delayed} = 0.42, t(46) = -29.92, 95\%CI = [0.38, 0.46], p < 10^{-15}$).

Using all models (Baseline Model, Differential Learning Model, Differential Weight Model and Differential Learning and Weight Model), the model with same learning rates and equal decision weights provides a better fit for 38% subjects, the model with differential learning rate and equal weights provides a better fit for 8% subjects, the model with equal learning rates and different weights provides a better fit for 47% of subjects, while the model with both different learning rates and different weights provides a better fit for 6% subjects.

Gaze Bias

On the whole, subjects did not have a tendency to look more at the immediate reward compared to the delayed reward. In a mixed-effects regression of relative dwell proportion on immediate vs. delayed, the effect (i.e., intercept) was not significant when controlling for the size and type (ascending vs. descending) of reward ($\beta = 0.002, 95\%CI = [-0.05, 0.06], p = .929$; Table S6), and also not significant when controlling for the predicted values and prediction errors ($\beta = 0.03, 95\%CI = [-0.04, 0.10], p = .444$; Table S7). The position of the immediate feedback at the bottom of the screen did not have a significant effect ($\beta_{ImmediateBottom} = -0.05, 95\%CI = [-0.15, 0.06], p = .381$; Table S6).

Unlike in the online eye-tracking experiment, subjects were more likely to fixate first to the immediate reward compared to the delayed reward. In a mixed-effects logistic regression of first fixation location (immediate vs. delayed) the effect (i.e., intercept) was significant when controlling for the size and type of reward ($\beta = 0.98, 95\%CI = [0.52, 1.45], p < 10^{-4}$; Table S6) and when controlling for the predicted values and prediction errors ($\beta = 0.67, 95\%CI = [0.14, 1.21], p = .014$; Table S7). However, position of the immediate feedback at the bottom of the screen made a significant difference, with subject being more likely to fixate first on the delayed reward

($\beta = 1.85, 95\%CI = [1.35, 2.35], p < 10^{-12}$; $\beta_{ImmediateBottom} = -1.91, 95\%CI = [-2.61, -1.22], p < 10^{-7}$; Table S7).

The learning bias was significantly correlated with the first fixation bias but not with the dwell proportion bias. However, this was not true when using the behavioral bias calculated from the regressions (Dwell Proportion: $\rho(45) = -.13, p = .4$; First Fixation: $\rho(45) = .01, p = .92$) but only when using the learning bias calculated from the RL model (Dwell Proportion: $\rho(45) = .29, p = .05$; First Fixation: $\rho(45) = .23, p = .12$; Fig. S21A,B).

Declarative Memory

At the end of the study, subjects completed a memory survey. We asked them to rank the stimuli in terms of total reward, to compare the immediate and delayed reward by indicating whether each stimulus was ascending or descending, and to estimate the average immediate reward and the average delayed reward for each stimulus.

Being worse at ranking stimuli was associated with lower accuracy similar to Study 2, but unlike in Study 2 where this was not significantly associated with the immediacy bias, here worse ranking memory was associated with a smaller immediacy bias. Similar to Study 2, subjects with a worse memory for whether a stimulus had a higher immediate or delayed reward were more accurate on incongruent choice sets and had a smaller behavioral bias. Almost all types of declarative memory errors, except for the point estimation of average delayed rewards, were associated with lower behavioral bias. In other words, remembering correctly leads to a larger, not smaller behavioral bias.

The error in the ranking of stimuli was positively related to the overall choice error rate ($\beta = 0.03, 95\%CI = [0.02, 0.05], p < 10^{-3}$; Table S9) and the error rate on the incongruent choice sets, but not significantly ($\beta = 0.03, 95\%CI = [-0.01, 0.08], p = .112$; Table S9). Moreover, this ranking error was significantly negatively related to the immediacy bias ($\beta_{Ranking} = -0.19, 95\%CI = [-0.36, -0.03], p = .024$; Table S10). This was not significant when measuring bias using the RL model ($\beta_{Ranking} = -0.003, 95\%CI = [-0.03, 0.02], p = .828$; Table S10). Namely, subjects with worse ranking memory had a lower immediacy bias.

The error rate in the ascending vs. descending survey was negatively related with the error rate on the incongruent choice sets ($\beta = -0.11, 95\%CI = [-0.18, -0.04], p = .004$; Table S9). Moreover, this memory error was significantly negatively related to the immediacy bias ($\beta_{Comparison} = -0.28, 95\%CI = [-0.56, -0.01], p = .046$; Table S10). Therefore, subjects with worse memory had again a lower immediacy bias. This was also true based on the bias measured using the RL model ($\beta_{Comparison} = -0.05, 95\%CI = [-0.096, -0.005], p = .031$; Table S10).

The estimation error for the delayed reward was positively related to the error rate on incongruent choice sets ($\beta = 0.15, 95\%CI = [0.02, 0.28], p = .021$; Table S9). Moreover, higher error in estimating points for the the immediate reward, but not delayed reward, was negatively associated with the immediacy bias ($\beta_{PointsImmediate} = -0.61, 95\%CI = [-1.16, -0.05], p = .032$; $\beta_{PointsDelayed} = 0.08, 95\%CI = [-0.42, 0.59], p = .74$; Table S10). This was not significant when measuring bias using the RL model ($\beta_{PointsImmediate} = -0.01, 95\%CI = [-0.10, 0.08], p = .822$; $\beta_{PointsDelayed} = 0.03, 95\%CI = [-0.06, 0.11], p = .538$; Table S10).

Working Memory

At the end of the study, subjects also completed a visual working memory task, namely a change localization task (*81*). Six colored squares appeared on the screen simultaneously. At the end of the trial, subjects saw the same six squares in their original locations, but one square had changed color. Subjects identified the changed square by pressing the corresponding number on the keyboard. During the test phase, each square was labeled with a number to facilitate response selection.

We regressed behavioral and learning biases on the accuracy for the working memory task. Accuracy was not significantly related to the behavioral or learning bias (Regression-based: $\beta = 0.65[-1.93, 3.22], p = .615$, RL-based: $\beta = 0.32[-0.04, 0.68], p = .081$; Table S11).

Discounting Preferences

We then regressed the intertemporal indifference point for all 3 time scales on the behavioral and learning-rate biases. There was no significant association between impatience and the behavioral bias (Regression-based: $\beta_{Today-1Month} = -2[-5.25, 1.24], p = .220$, RL-based: $\beta_{Today-1Month} = -10.34[-33.10, 12.42], p = .365$, Regression-based: $\beta_{Today-6Months} = -1.94[-5.55, 1.66], p = .283$, RL-based: $\beta_{Today-6Months} = -11.13[-36.31, 14.05], p = .378$, Regression-based: $\beta_{1Month-6Months} = -1.48[-4.99, 2.04], p = .402$, RL-based: $\beta_{1Month-6Months} =$

$-19.45[-43.40, 4.50], p = .109$; Table S13). Unlike in Study 1, the higher the impatience, the lower the immediacy bias, but no effects were significant.

Additional References

1. Greiner, B. Subject pool recruitment procedures: organizing experiments with ORSEE. *J Econ Sci Assoc* **1**, 114–125 (2015).
2. Falk, A., Becker, A., Dohmen, T., Huffman, D. & Sunde, U. The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *IZA Discussion Paper* **9674** 1–66 (2016).
3. Harel, J., Koch, C. & Perona, P. Graph-Based Visual Saliency. in *Advances in Neural Information Processing Systems 19* (eds. Schölkopf, B., Platt, J. & Hofmann, T.) 545–552 (The MIT Press, 2007).
4. Zhao, C., Vogel, E. & Awh, E. Change localization: A highly reliable and sensitive measure of capacity in visual working memory. *Atten Percept Psychophys* **85**, 1681–1694 (2023).

Supplementary Figures

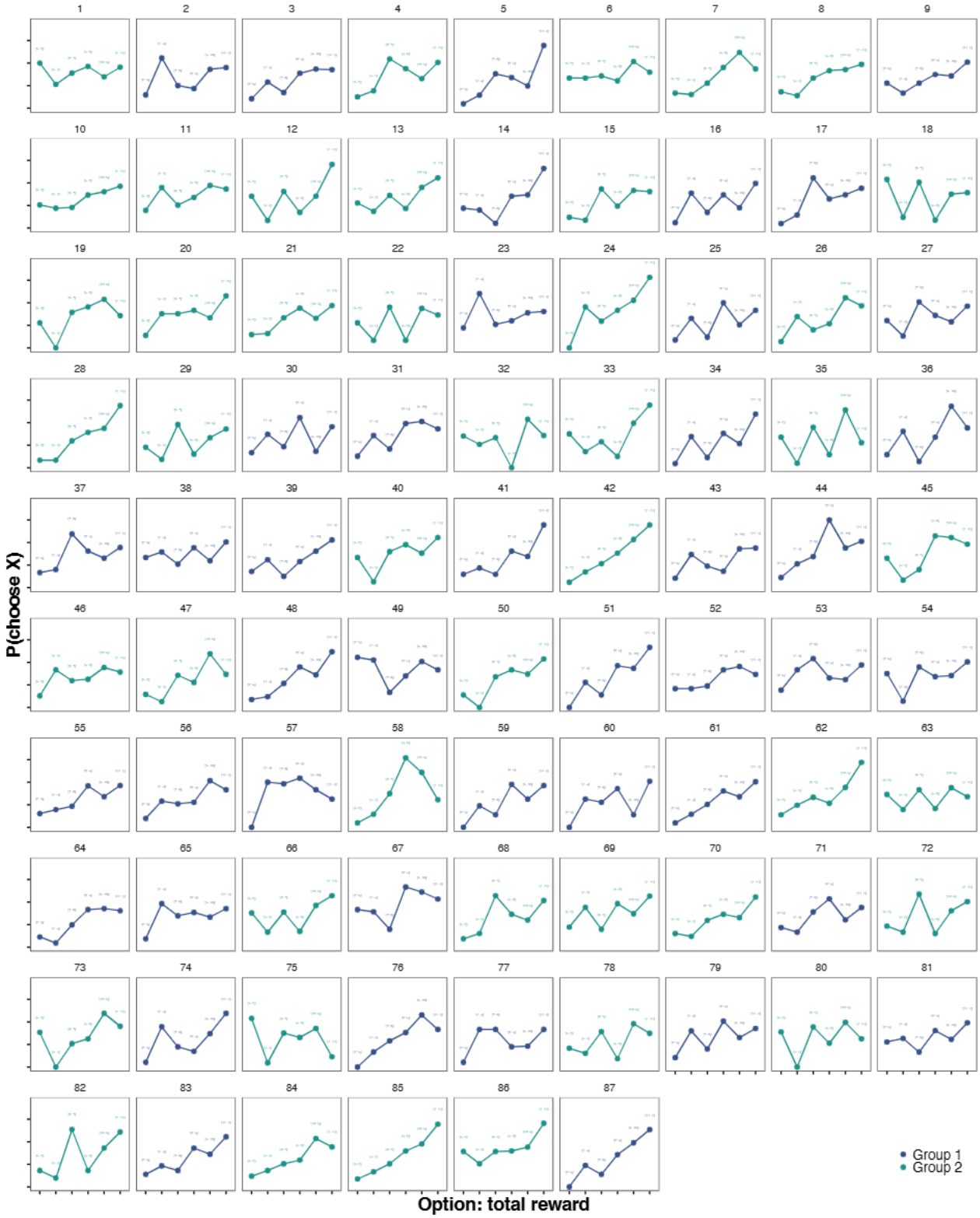


Figure S1. Behavioral bias at the subject level for Study 1 Colors. Probability of choosing an option as a function of the option's total reward for Study 1 Colors.

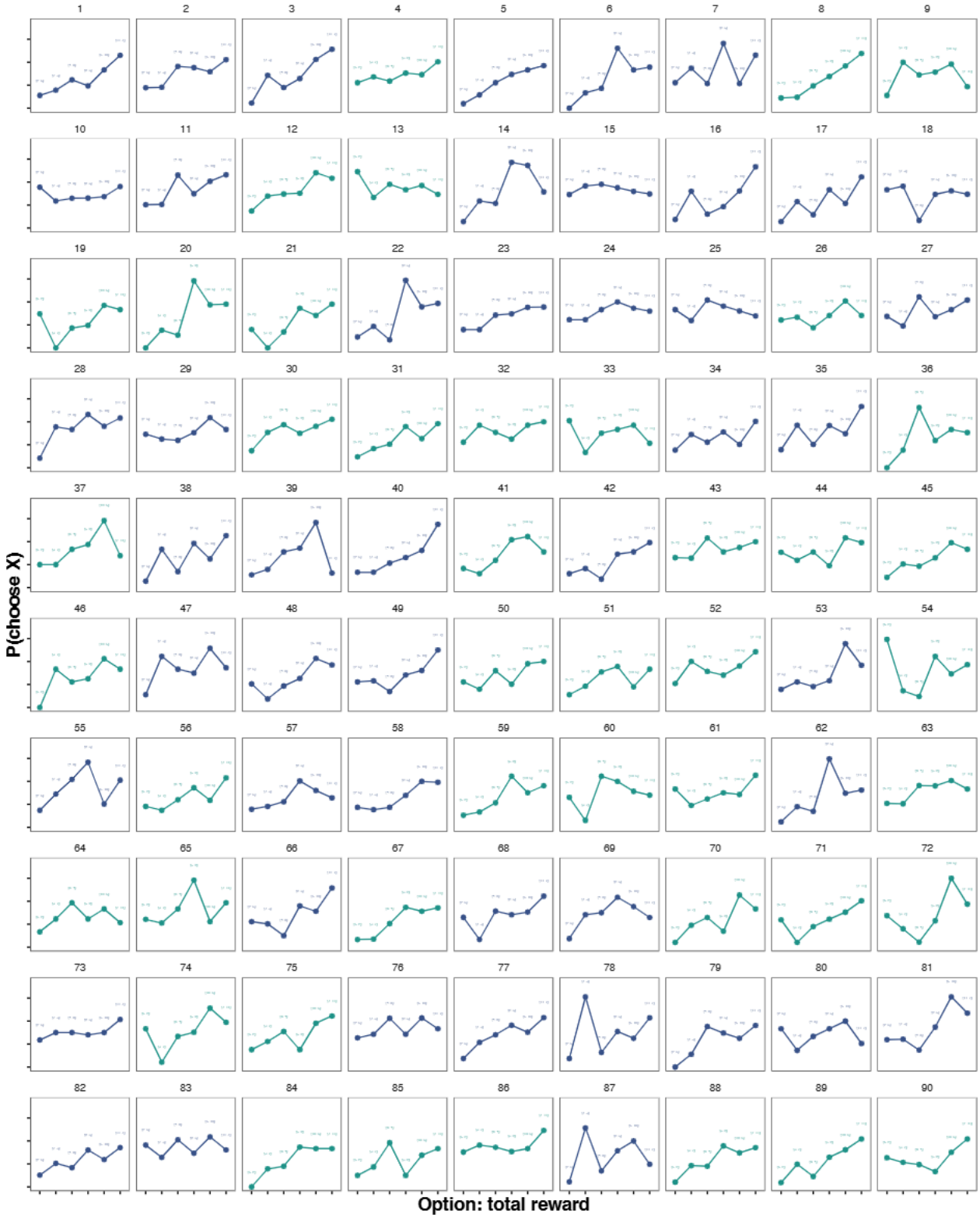


Figure S2. Behavioral bias at the subject level for Study 2 Patterns. Probability of choosing an option as a function of the option's total reward for Study 2 Patterns.

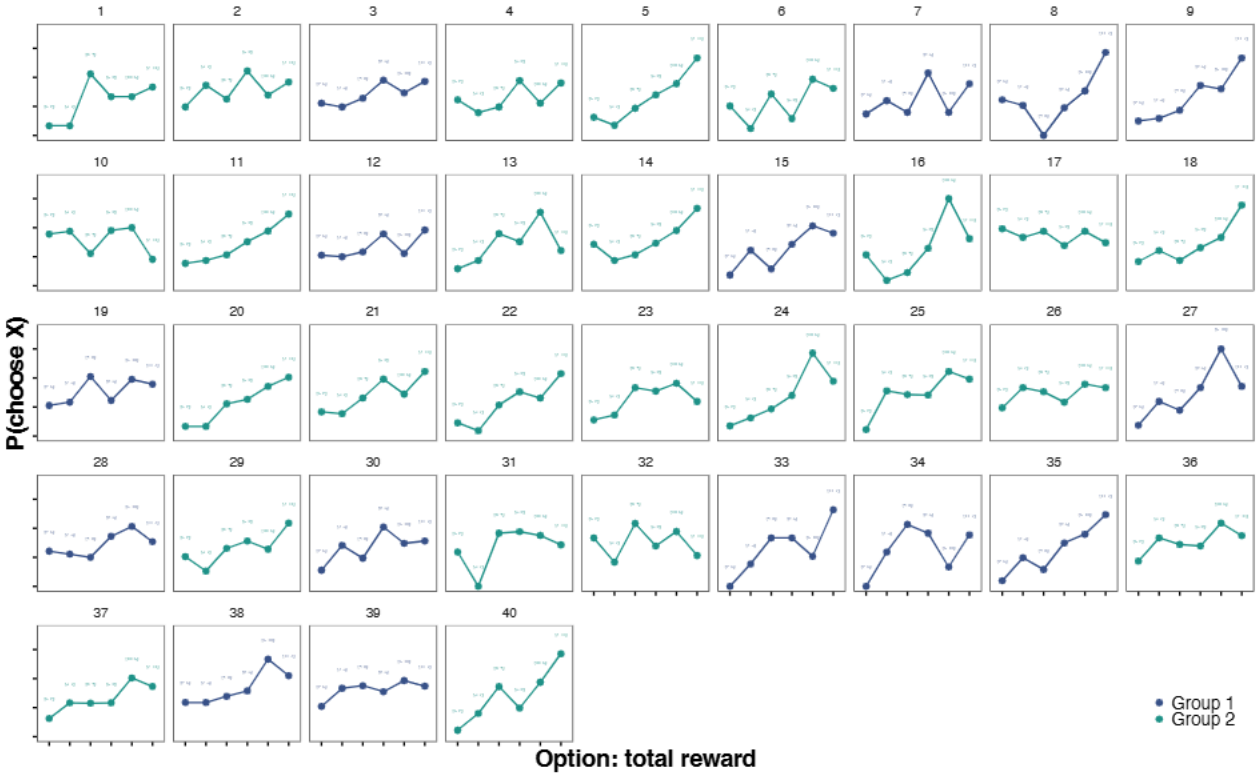


Figure S3. Behavioral bias at the subject level for Study 3 Patterns Reversed Feedback Position. Probability of choosing an option as a function of the option's total reward for Study 3 Patterns Reversed Feedback Position.

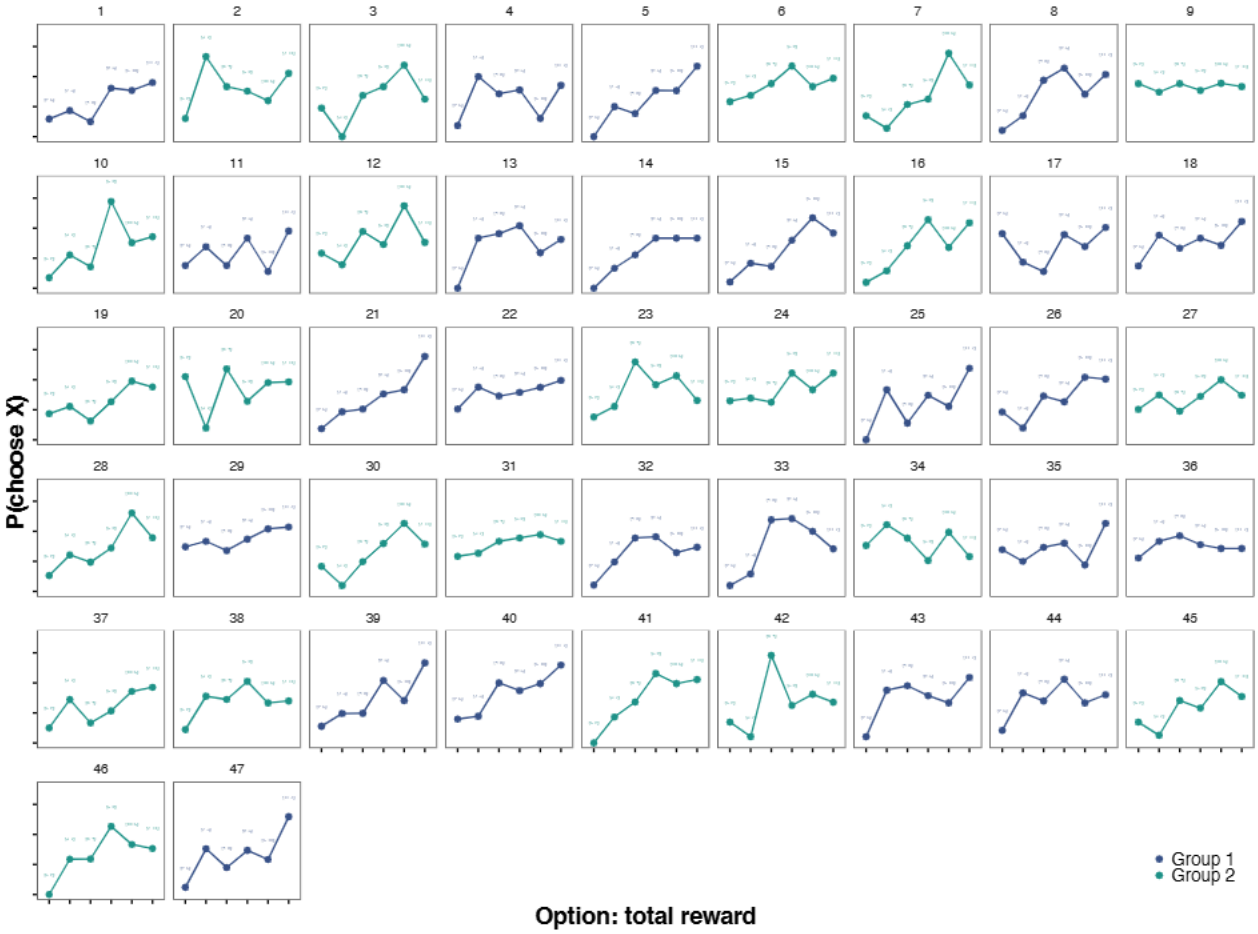


Figure S4. Behavioral bias at the subject level for Study 4 In-lab Eye-tracking. Probability of choosing an option as a function of the option's total reward for Study 4 In-lab Eye-tracking.

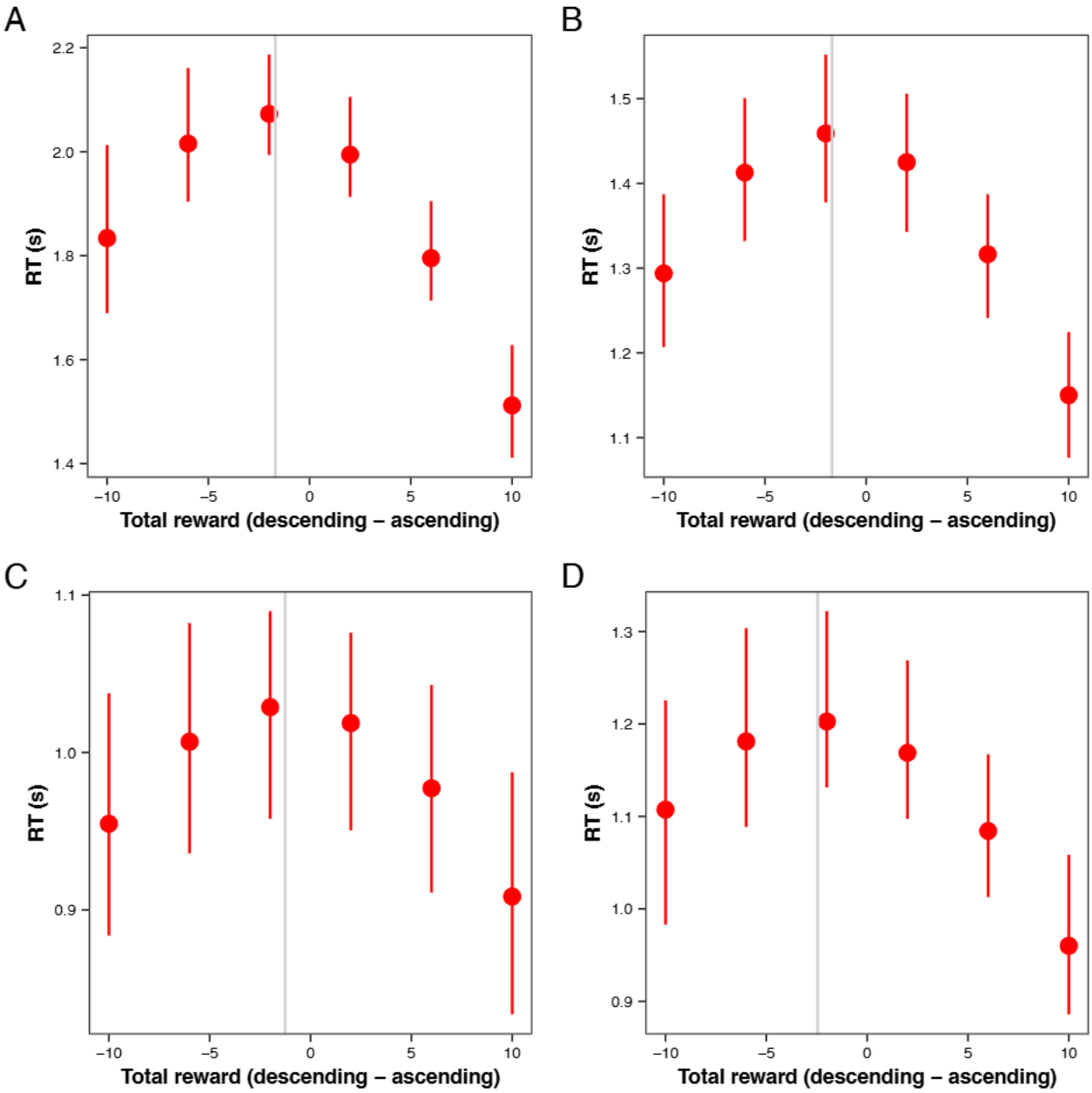


Figure S5. RT indifference point. **(A,B,C)** Predicted RT in seconds from a mixed effects quadratic regression using difference in total underlying reward between right and left stimulus only for choice sets where one option is descending and the other is ascending. The red bars represent 95% confidence intervals. The gray line represents the predicted maximum RT. The indifference point occurs when the descending option is slightly worse than the ascending option. **(A)** Colors Task. **(B)** Patterns Task. **(A,B)** Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. **(C)** Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. **(D)** In-lab Eye-tracking.

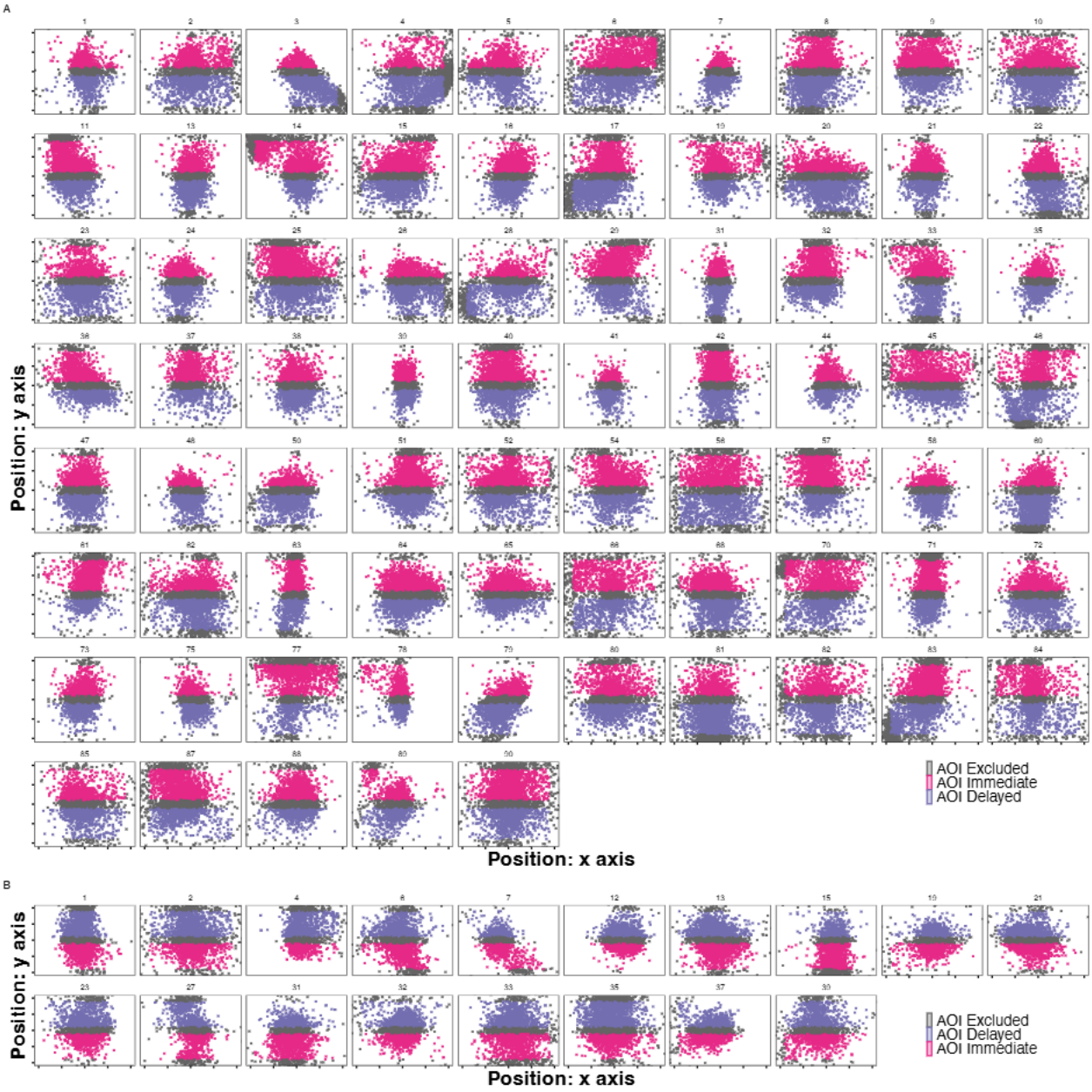


Figure S6. Attention to feedback screen for each subject for Studies 2 and 3 Patterns. Online Patterns Study. **(A)** Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. **(B)** Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. Each dot represents one fixation.

Figure S7. Attention to feedback screen for each subject for Study 4 In-lab Eye-tracking. In-lab Eye-tracking Study. Each dot represents one fixation.

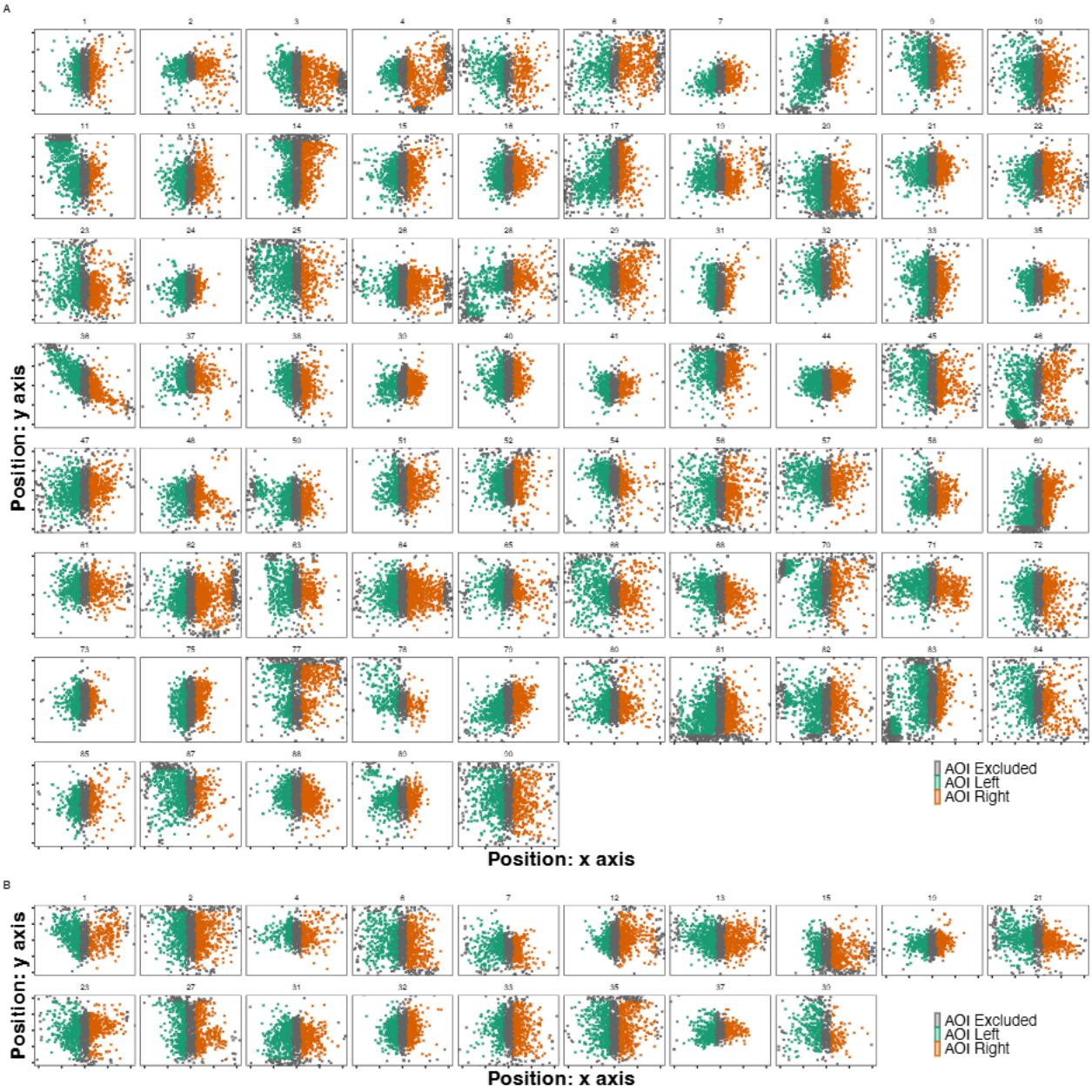


Figure S8. Attention to the choice screen for each subject for Studies 2 and 3 Patterns. Online Patterns Study. **(A)** Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. **(B)** Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. Each dot represents one fixation.

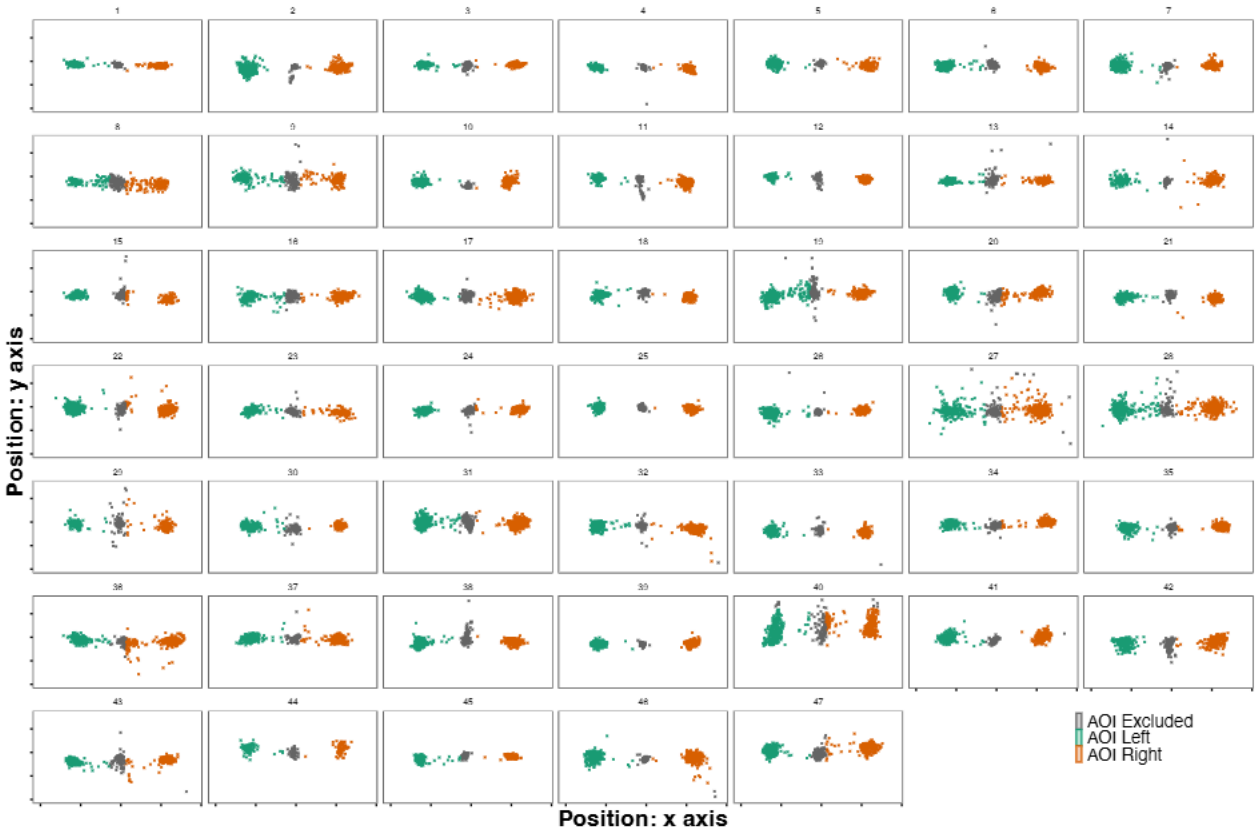


Figure S9. Attention to the choice screen for each subject for Study 4 In-lab Eye-tracking. In-lab Eye-tracking Study. Each dot represents one fixation.

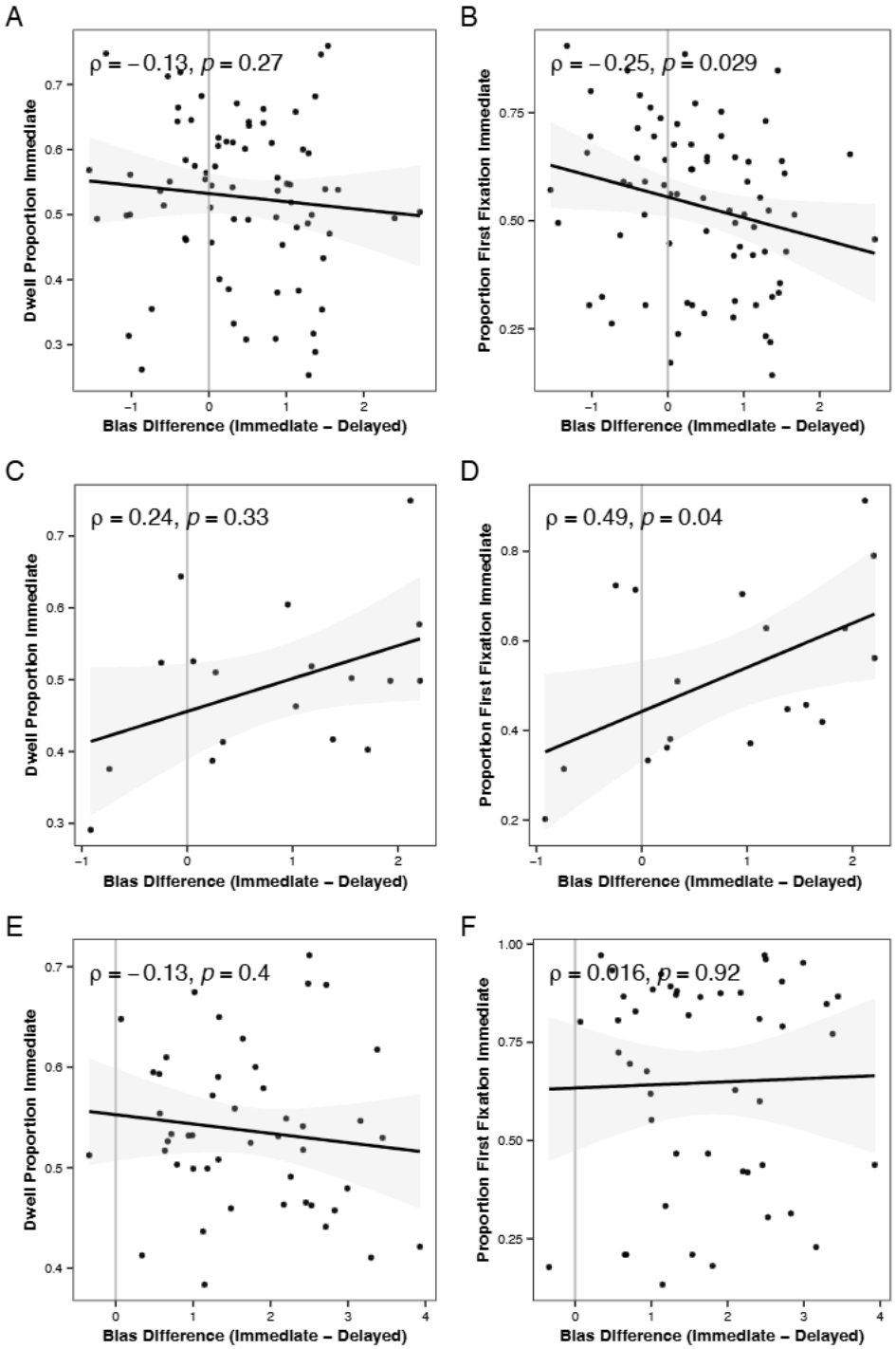


Figure S10. Behavioral bias and attention. **(A,C,E)** Correlation between behavioral bias difference (the difference between coefficients in the experienced average immediate reward and delayed reward from mixed effects logistic regression of choosing the left stimulus) and average proportion of dwell time to the immediate reward across trials. **(B,D,F)** Correlation between behavioral bias difference and average proportion of first fixation to immediate rewards across trials. **(A,B)** Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. **(C,D)** Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. **(E,F)** In-lab Eye-tracking Study. Each dot represents a subject. The black line represents the best fitting linear regression line. The gray band represents the 95% CI.

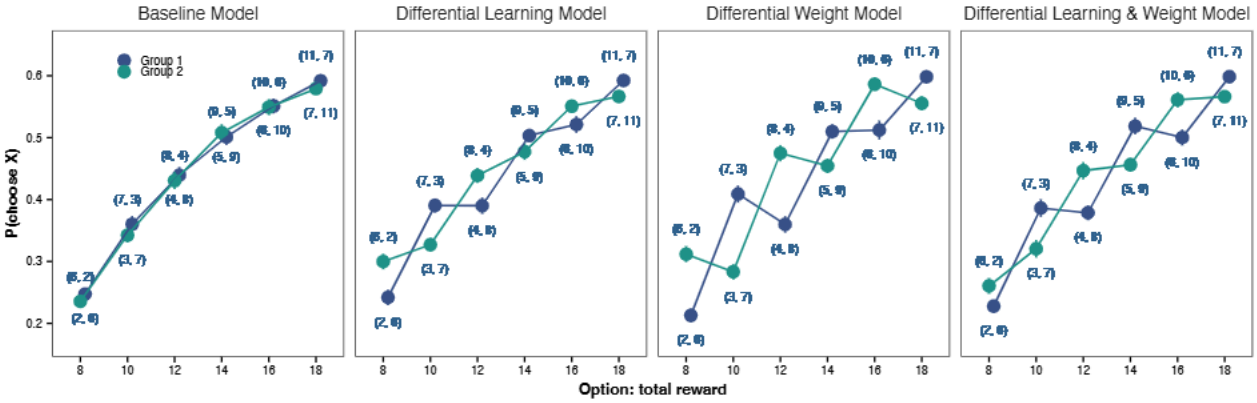


Figure S11. Predicted aggregate behavioral bias. Predicted probability of choosing a stimulus given it is in the choice set as a function of the total reward of the stimulus for each condition. (**A**) For Baseline Model (same learning rate for immediate and delayed rewards). (**B**) For Differential Learning Model (different learning rates for immediate and delayed rewards). (**C**) Differential Weight Model (same learning rates for immediate and delayed rewards, weight on delayed reward). (**D**) For Differential Learning and Weight Model (different learning rates for immediate and delayed rewards and weight on delayed reward). The dots represent averages across the 100 simulated datasets for each subject in the Colors Task (Study 1) and Patterns Task (Study 2).

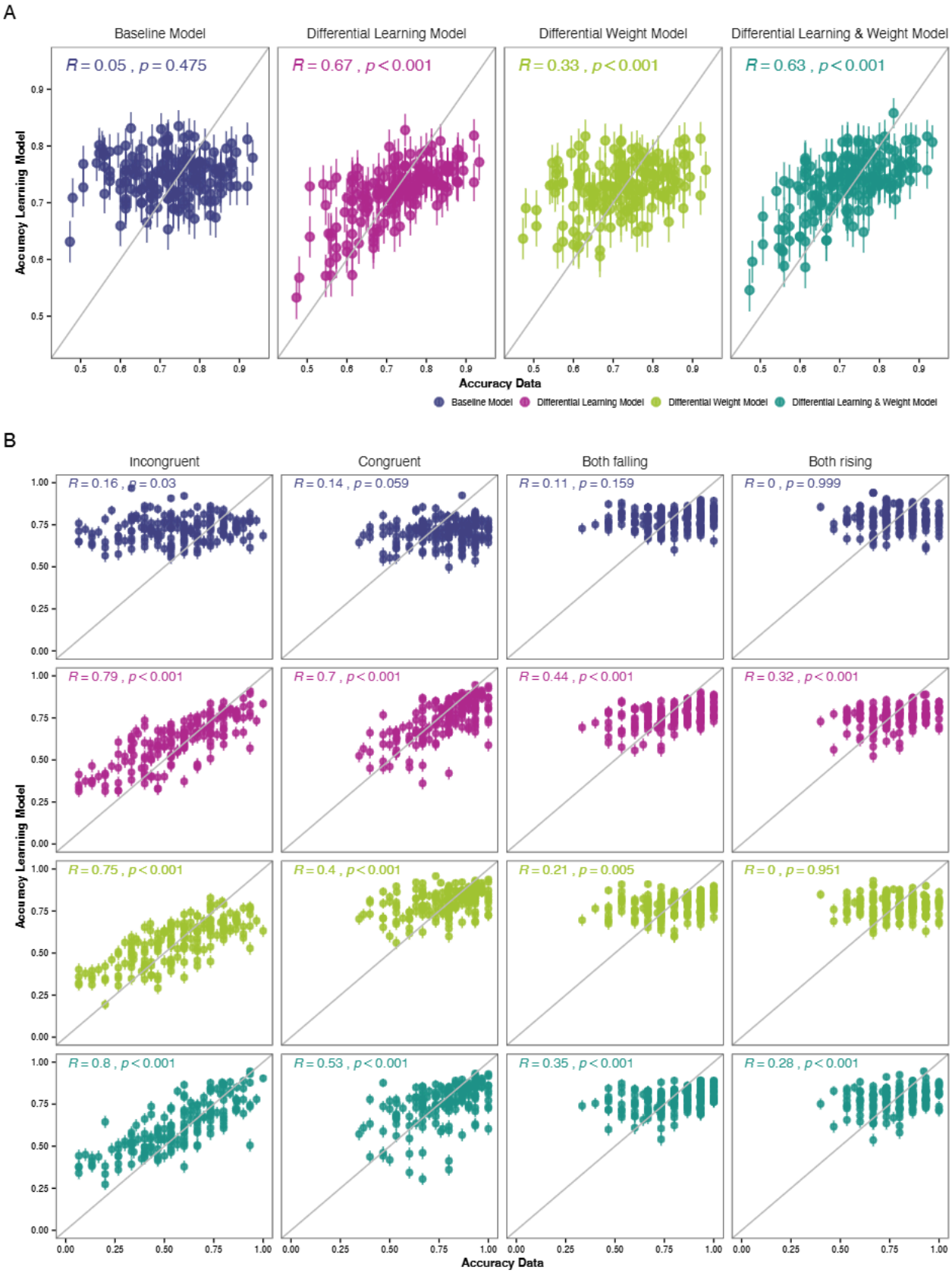


Figure S12. Predicted and observed accuracy at the subject level. **(A)** Accuracy in data versus model. **(B)** Accuracy

in the data versus model. Incongruent choice sets: worse option descending, better option ascending, Congruent choice sets: worse option ascending, better option descending. The dots represent subjects. The bars represent standard errors of the 100 simulated datasets for each subject in the Colors Task (Study 1) and Patterns Task (Study 2).

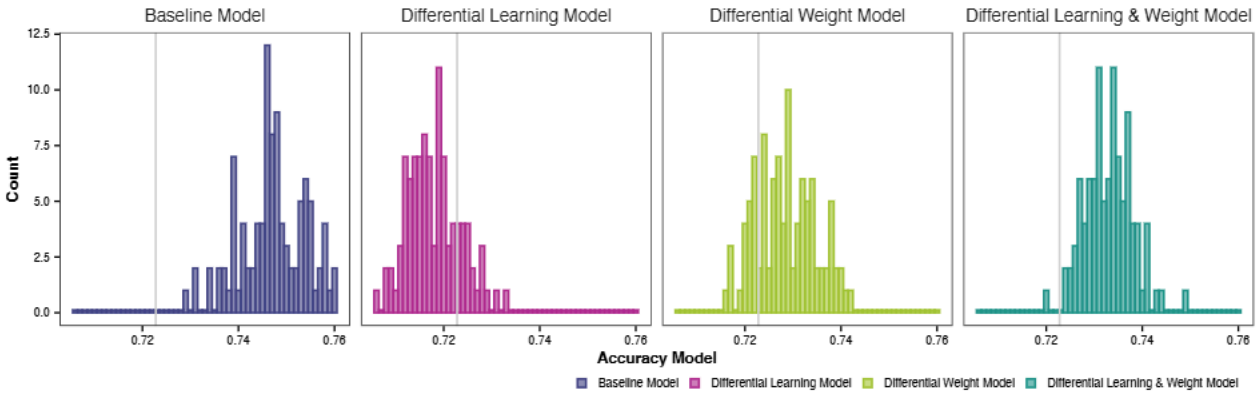


Figure S13. Observed and predicted experiment accuracy. Histogram of model predicted mean accuracy across subjects in the Colors Task (Study 1) and Patterns Task (Study 2) for the 100 simulated datasets for each model. The gray line represents the mean accuracy across subjects in the data.

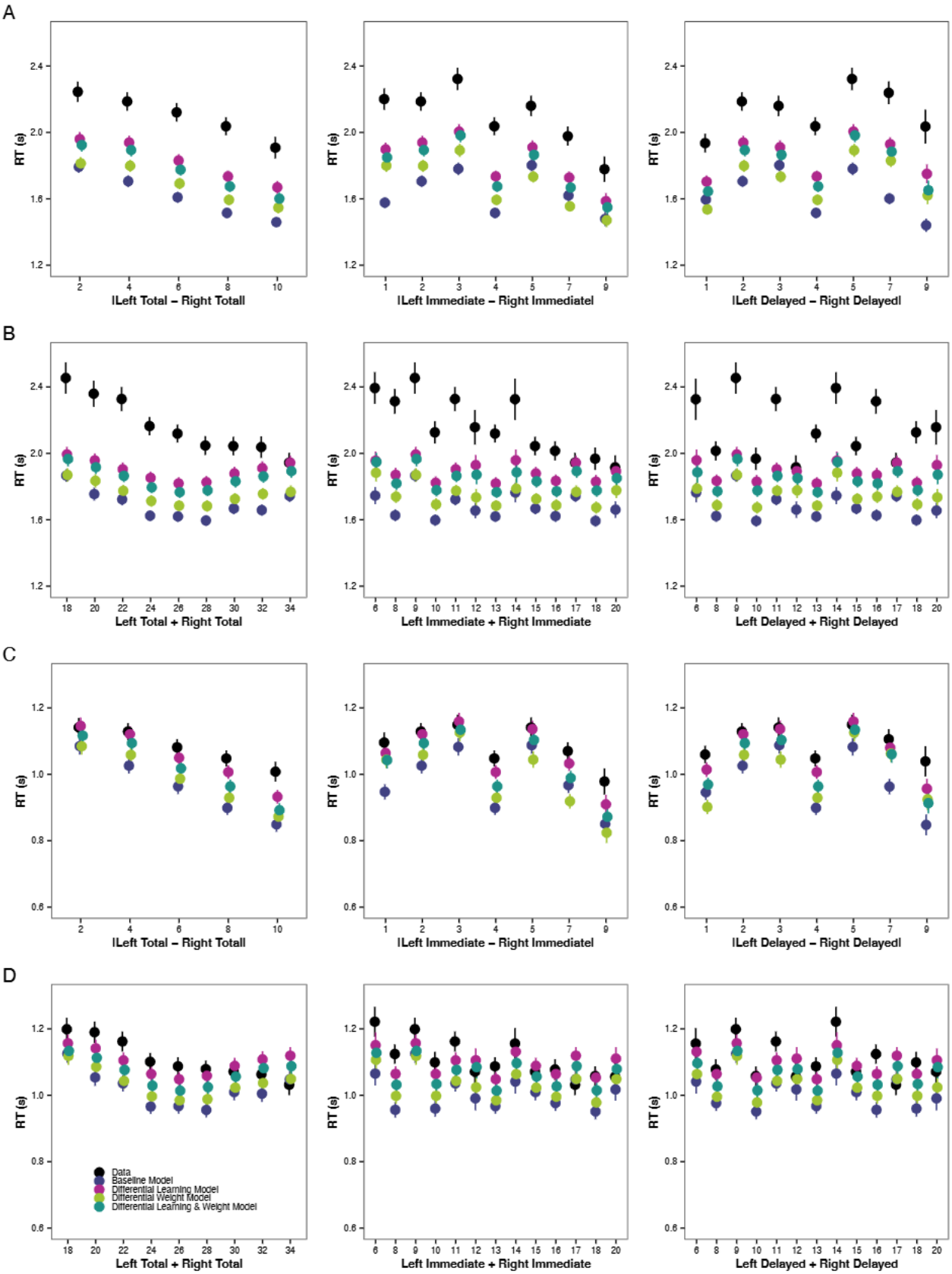
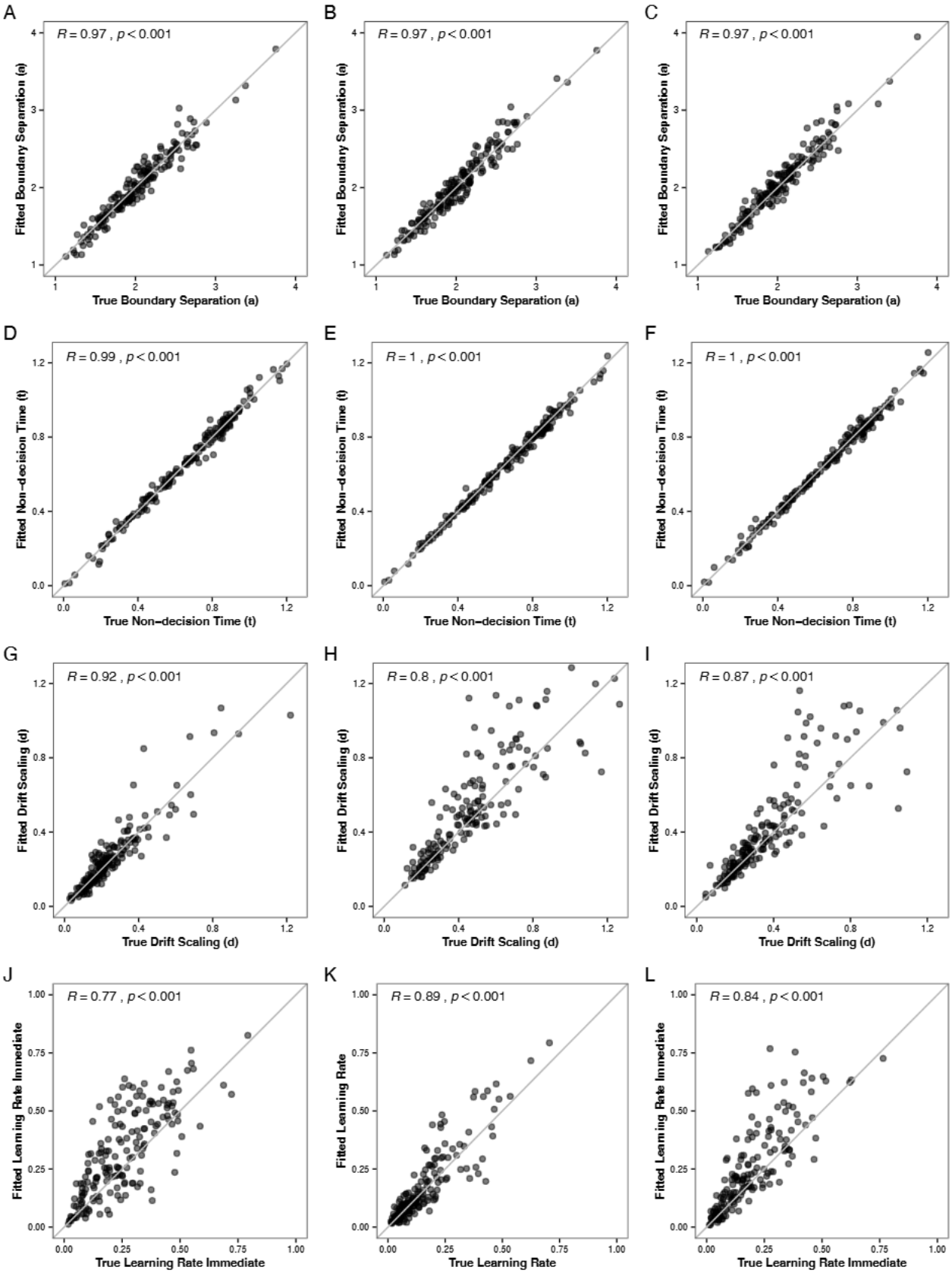


Figure S14. Observed and predicted value difference (VD) and overall value (OV) effects on RT. (A,B) Colors

Task (Study 1). (**C,D**) Patterns Task (Study 2). (**A,C**) VD. (**B,D**) OV. The model data was based on the 100 simulated datasets for each subject in the Colors Task (Study 1) and Patterns Task (Study 2).


A
R = 0.97 , p < 0.001
Fitted Boundary Separation (a)
True Boundary Separation (a)
B
R = 0.97 , p < 0.001
Fitted Boundary Separation (a)
True Boundary Separation (a)
C
R = 0.97 , p < 0.001
Fitted Boundary Separation (a)
True Boundary Separation (a)
D
R = 0.99 , p < 0.001
Fitted Non−decision Time (t)
True Non−decision Time (t)
E
R = 1 , p < 0.001
Fitted Non−decision Time (t)
True Non−decision Time (t)
F
R = 1 , p < 0.001
Fitted Non−decision Time (t)
True Non−decision Time (t)
G
R = 0.92 , p < 0.001
Fitted Drift Scaling (d)
True Drift Scaling (d)
H
R = 0.8 , p < 0.001
Fitted Drift Scaling (d)
True Drift Scaling (d)
I
R = 0.87 , p < 0.001
Fitted Drift Scaling (d)
True Drift Scaling (d)
J
R = 0.77 , p < 0.001
Fitted Learning Rate Immediate
True Learning Rate Immediate
K
R = 0.89 , p < 0.001
Fitted Learning Rate
True Learning Rate
L
R = 0.84 , p < 0.001
Fitted Learning Rate Immediate
True Learning Rate Immediate

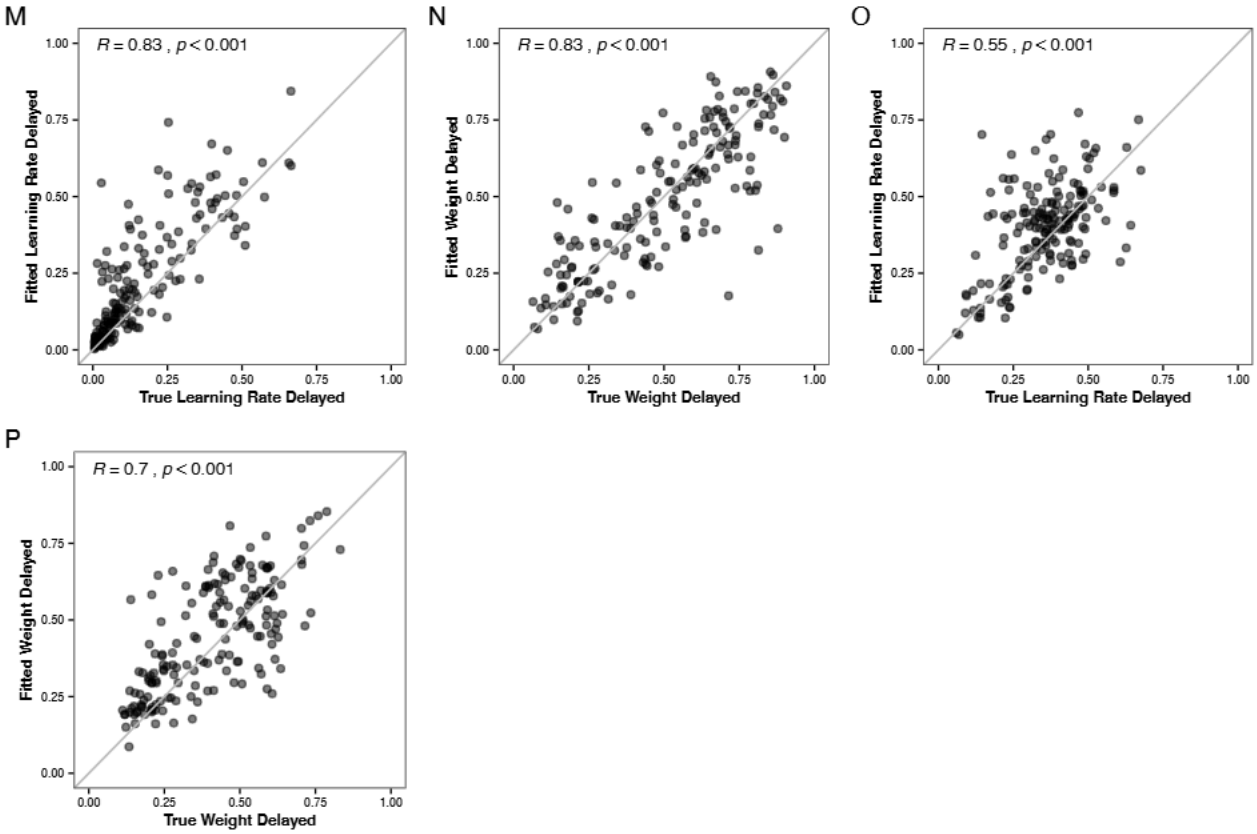


Figure S15. RL parameters recovery. Correlations between true parameter value used to generate the data and the mean posterior of the parameter values for the Differential Learning Model. Each dot represents a simulated subject. **(A,B,C)** Boundary separation. **(D,E,F)** Non-decision time. **(G,H,I)** Drift scaling parameter. **(K)** Learning rate parameter. **(J,L)** Immediate learning rate. **(M,O)** Delayed learning rate. **(N,P)** Weight parameter. **(A,D,G,J,M)** Differential Learning Model. **(B,E,H,K,N)** Differential Weight Model. **(C,F,I,L,O,P)** Differential Learning and Weight Model.

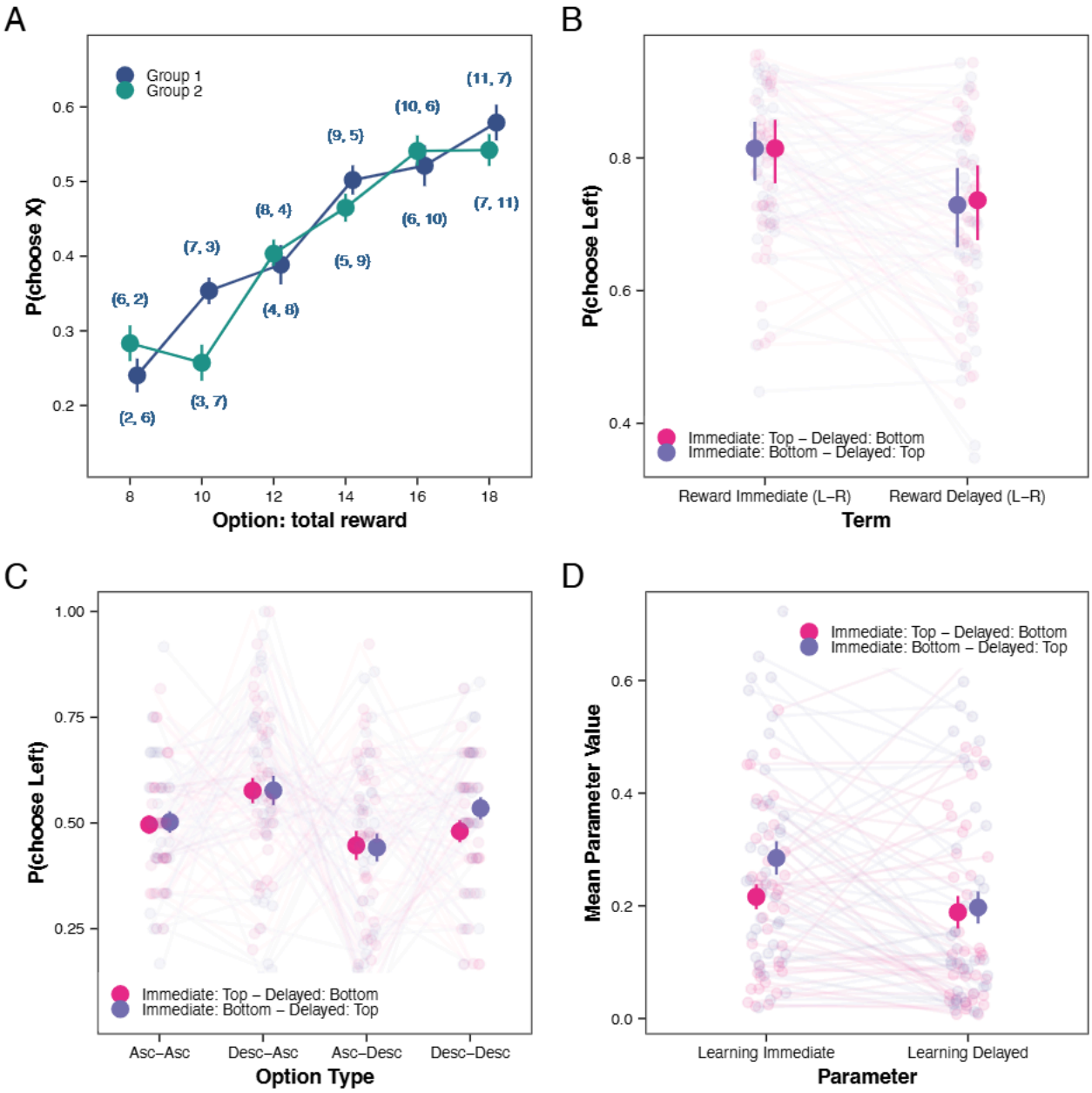


Figure S16. Behavioral bias for reversed feedback position. **(A)** Probability of choosing a stimulus given it is in the choice set as a function of the total reward of the stimulus. For the same total value of the option, the descending option is more likely to be chosen than the ascending option. **(B)** Probability of choosing the left stimulus as a function of the difference between the left and right option in the experienced immediate reward and delayed reward, based on a mixed-effects logistic regression. Dots represent subject level effects and bars represent standard errors of the fixed effects. **(C)** Probability of choosing the left stimulus as a function of whether the left option was descending or ascending. Dots represent subject level averages and bars represent standard errors across subjects. **(D)** Mean and standard errors of posteriors of immediate and delayed learning rates. Dots represent subject level learning rates for each task.

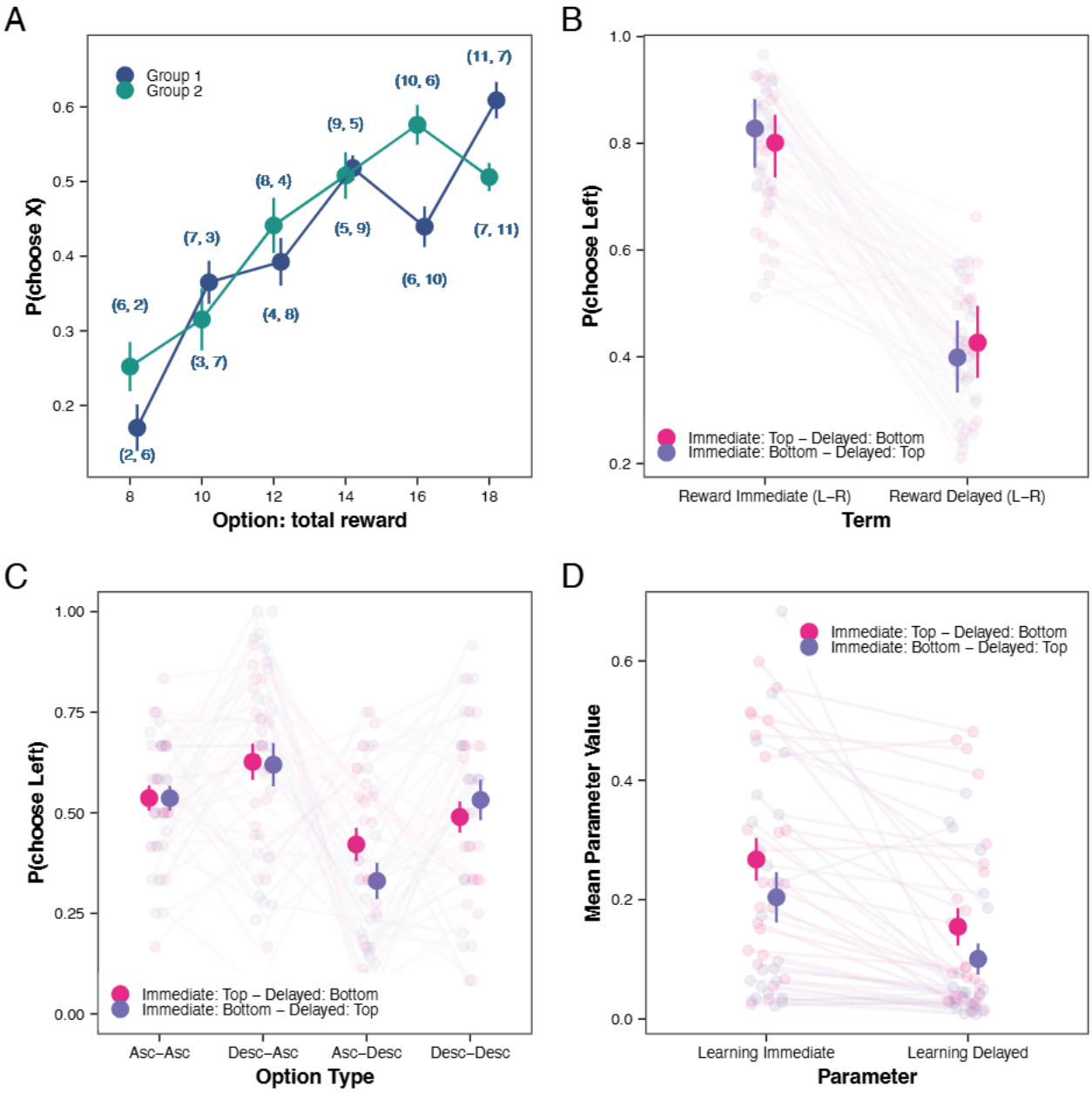


Figure S17. Behavioral bias for in-lab eye-tracking study. **(A)** Probability of choosing a stimulus given it is in the choice set as a function of the total reward of the stimulus. For the same total value of the option, the descending option is more likely to be chosen than the ascending option. **(B)** Probability of choosing the left stimulus as a function of the difference between the left and right option in the experienced immediate reward and delayed reward, based on a mixed-effects logistic regression. Dots represent subject level effects and bars represent standard errors of the fixed effects. **(C)** Probability of choosing the left stimulus as a function of whether the left option was descending or ascending. Dots represent subject level averages and bars represent standard errors across subjects. **(D)** Mean and standard errors of posteriors of immediate and delayed learning rates. Dots represent subject level learning rates for each task.

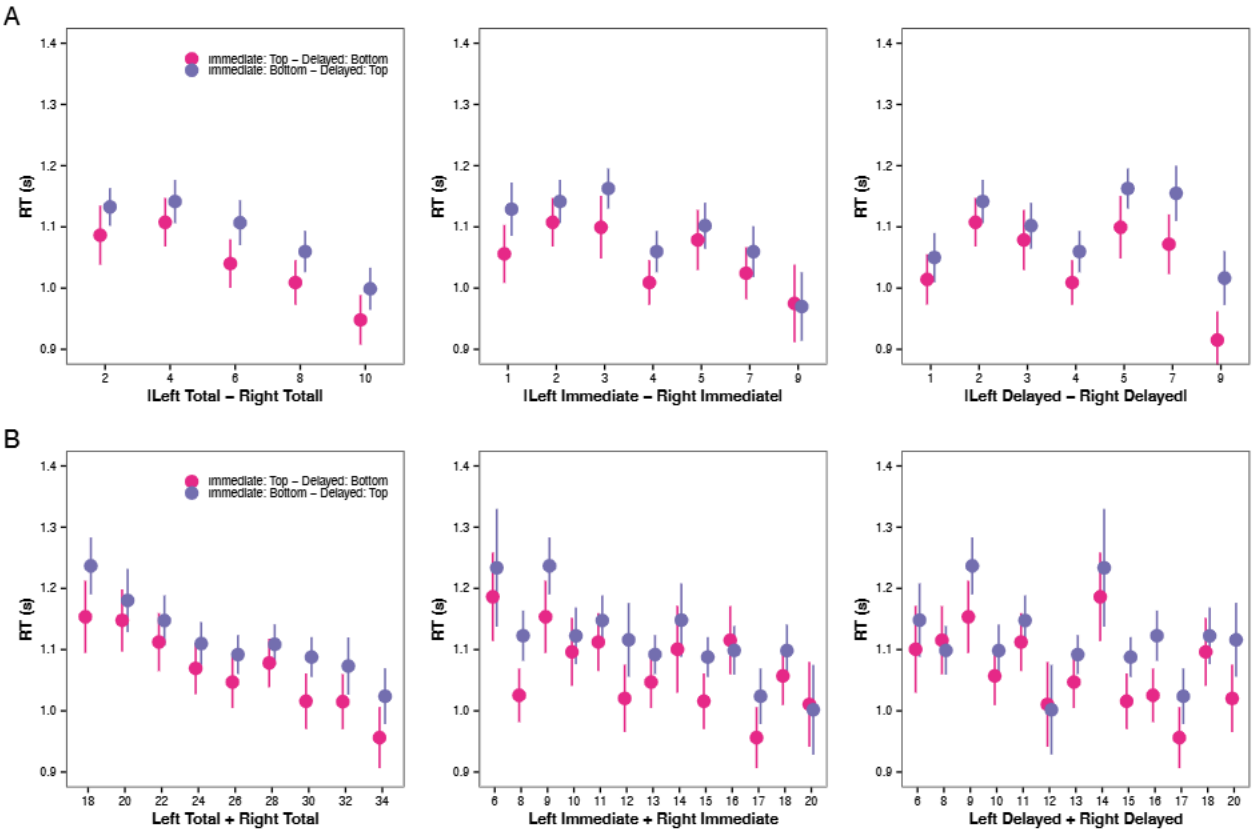


Figure S18. Value difference (VD) and overall value (OV) effects on response times (RT) for reversed feedback position condition. **(A)** Absolute value difference (|VD|) effects on RT for each Study. RT decreases with total |VD|. RT decreases with immediate |VD|. RT does not decrease with delayed |VD|. **(B)** Overall value (OV) effects on RT for each study. RT decreases with total OV. RT decreases with immediate OV. RT weakly decreases with delayed OV. Dots and bars represent mean and standard errors across subjects for each task.

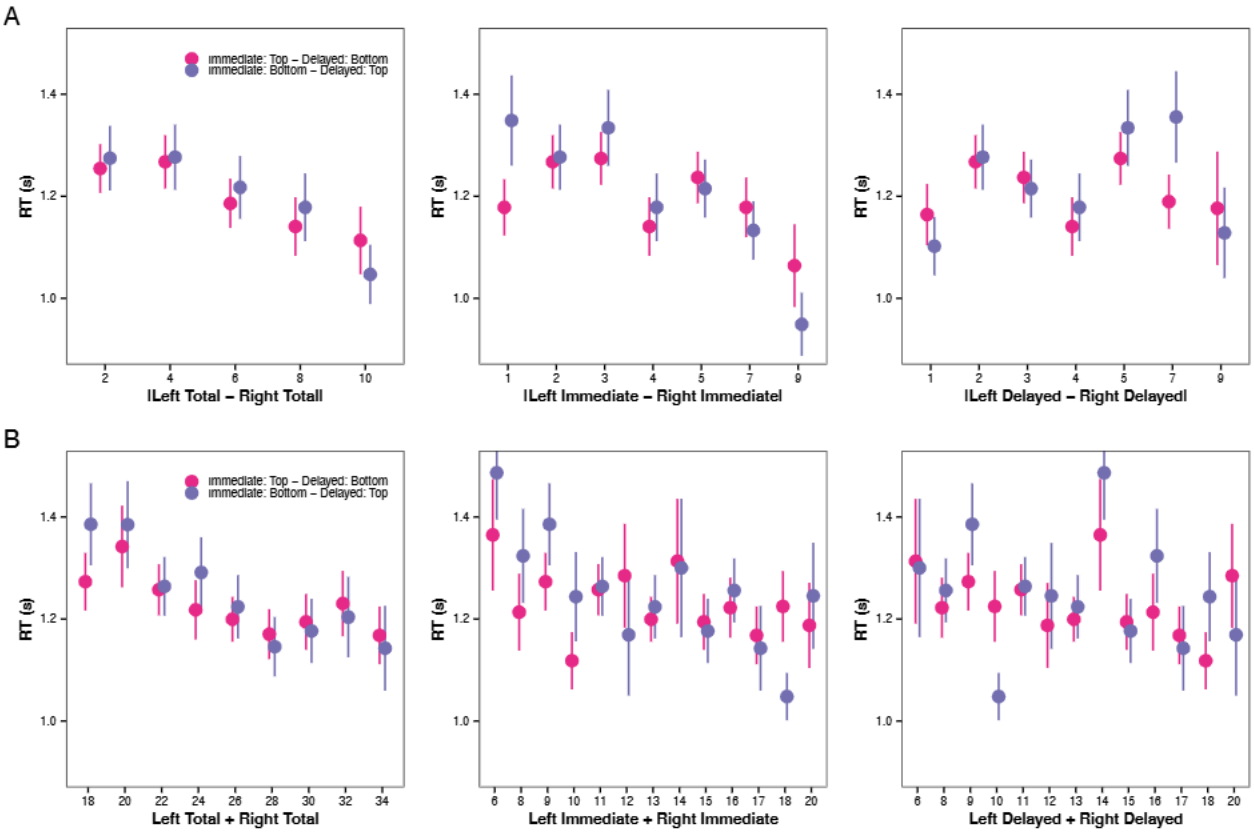


Figure S19. Value difference (VD) and overall value (OV) effects on response times (RT) for in-lab eye-tracking study. **(A)** Absolute value difference (|VD|) effects on RT for each Study. RT decreases with total |VD|. RT decreases with immediate |VD|. RT does not decrease with delayed |VD|. **(B)** Overall value (OV) effects on RT for each study. RT decreases with total OV. RT decreases with immediate OV. RT weakly decreases with delayed OV. Dots and bars represent mean and standard errors across subjects for each task.

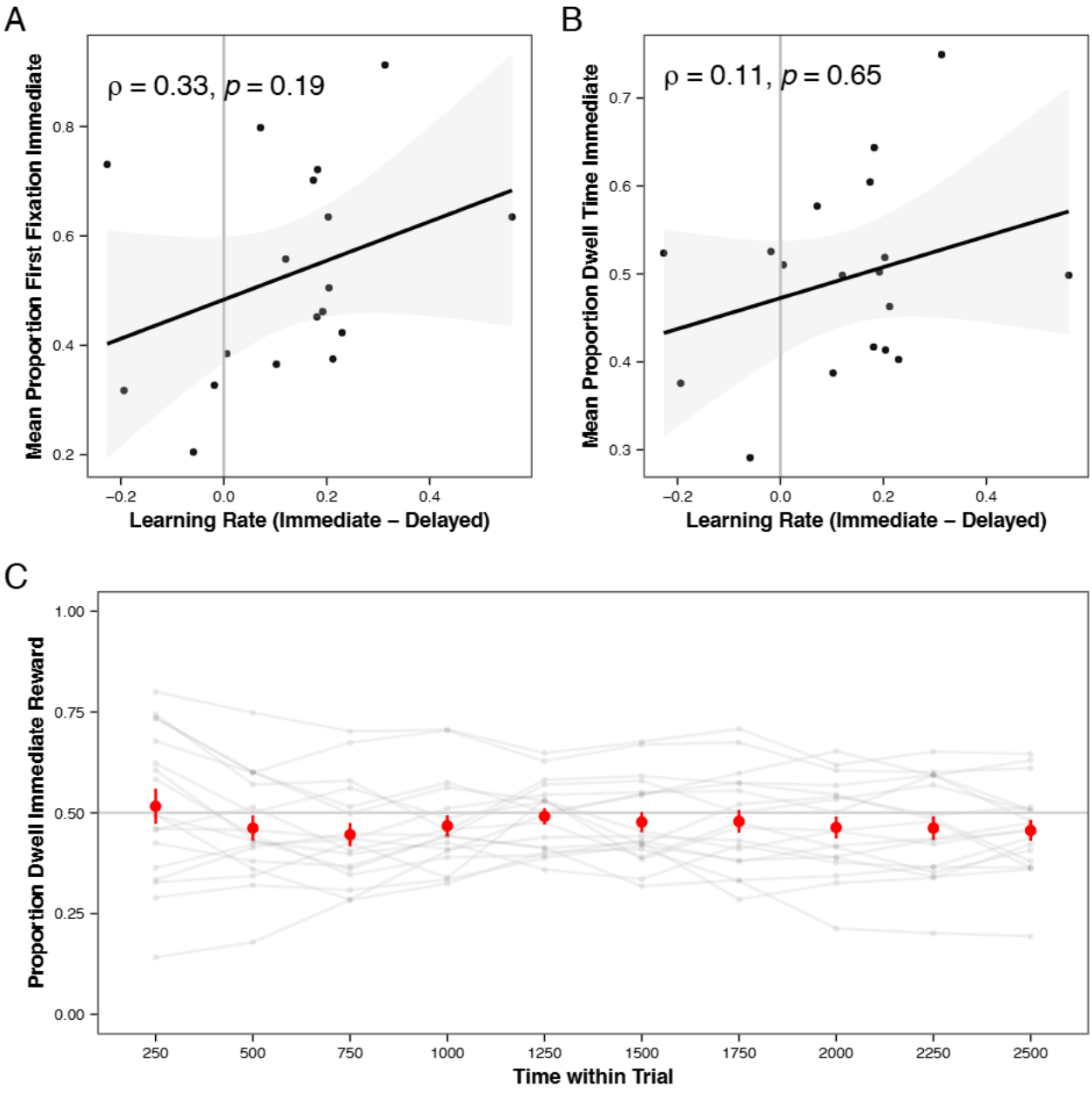
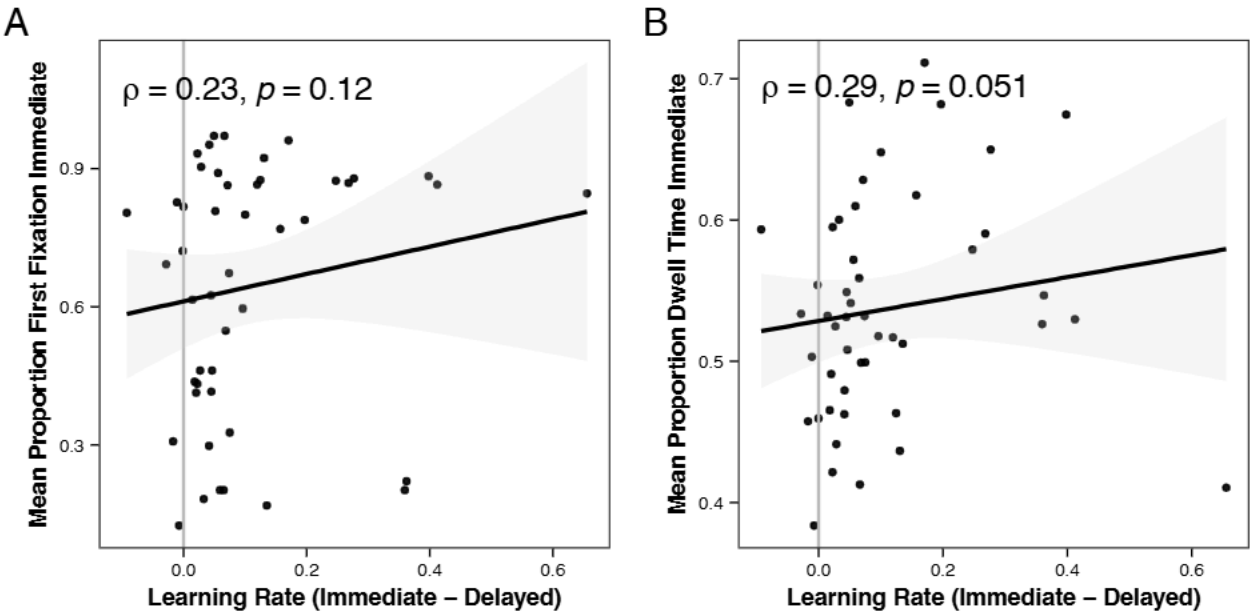
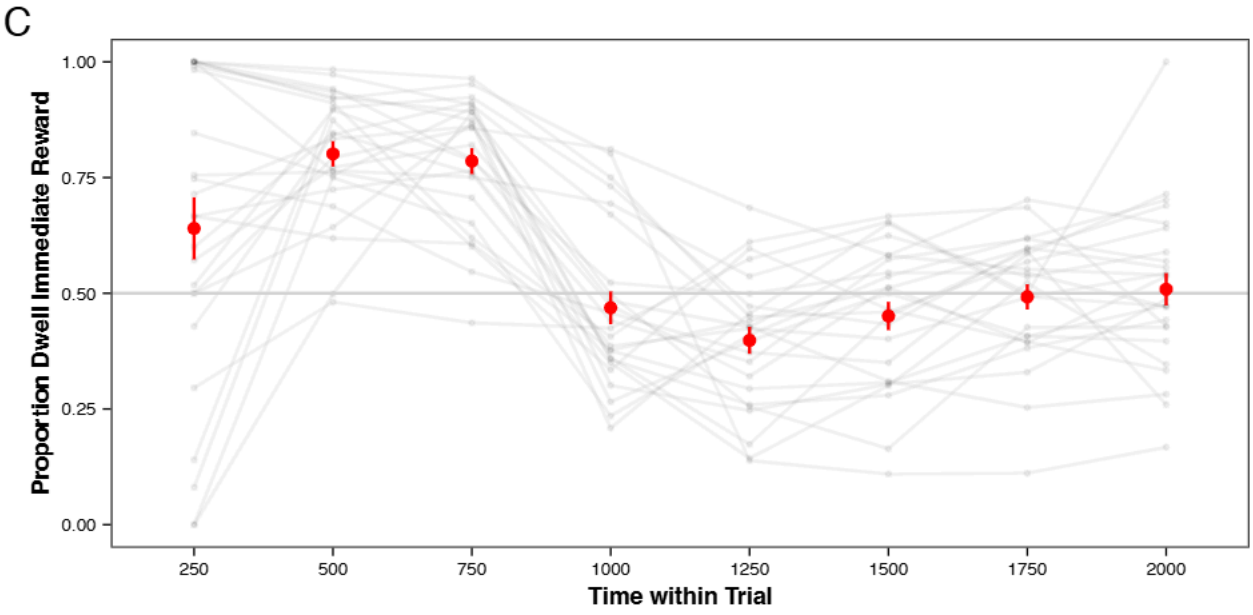


Figure S20. Attention and behavioral bias for reversed feedback position. **(A)** Correlation between difference in learning rate between immediate and delayed reward and proportion of first fixation to the immediate reward. **(B)** Correlation between difference in learning rate between immediate and delayed reward and mean dwell proportion immediate reward across trials. **(A,B)** Dots represent each subject. The black line represents the best fitting linear regression line. The gray band represents the 95% CI. **(C)** Mean and standard errors of dwell proportion to immediate reward within a trial for each time bin. Black dots represent each subject within a time bin.


A
ρ = 0.23, p = 0.12
Mean Proportion First Fixation Immediate
0.9
0.6
0.3
0.0
0.2
0.4
0.6
Learning Rate (Immediate – Delayed)
B
ρ = 0.29, p = 0.051
Mean Proportion Dwell Time Immediate
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
Learning Rate (Immediate – Delayed)


C
Proportion Dwell Immediate Reward
1.00
0.75
0.50
0.25
0.00
250
500
750
1000
1250
1500
1750
2000
Time within Trial

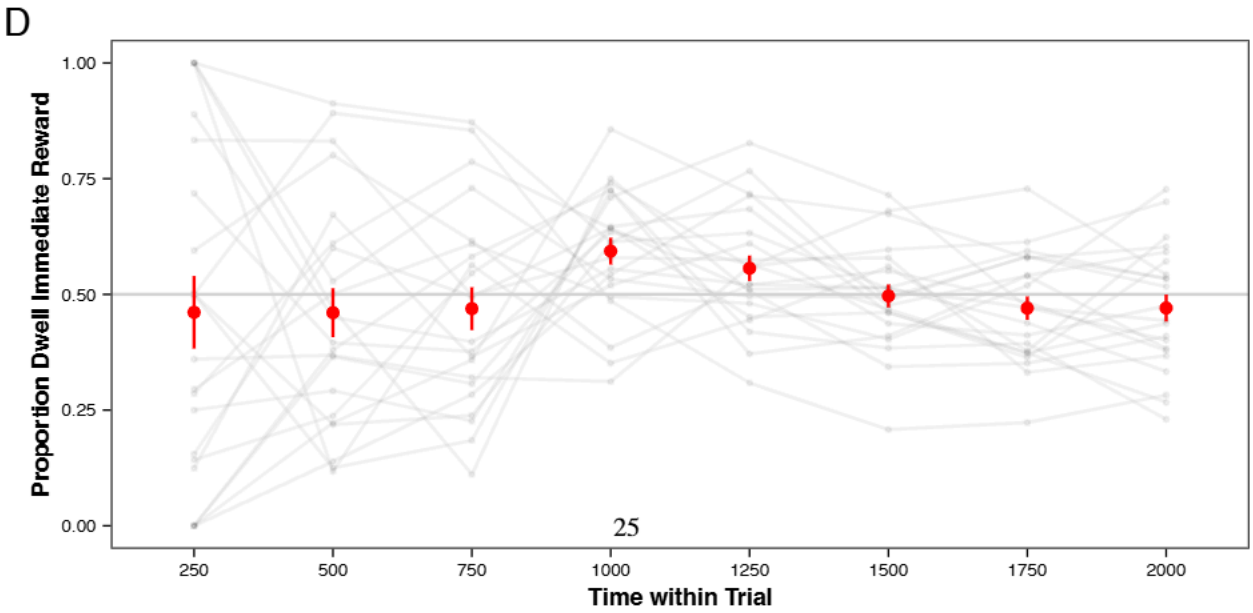

D
Proportion Dwell Immediate Reward
1.00
0.75
0.50
0.25
0.00
250
500
750
1000
1250
1500
1750
2000
Time within Trial

Figure S21. Attention and behavioral bias for in-lab eye-tracking data. **(A)** Correlation between difference in learning rate between immediate and delayed reward and proportion of first fixation to the immediate reward. **(B)** Correlation between difference in learning rate between immediate and delayed reward and mean dwell proportion immediate reward across trials. **(A,B)** Dots represent each subject. The black line represents the best fitting linear regression line. The gray band represents the 95% CI. **(C,D)** Mean and standard errors of dwell proportion to immediate reward within a trial for each time bin. **(C)** Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. **(D)** Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. Black dots represent each subject within a time bin.

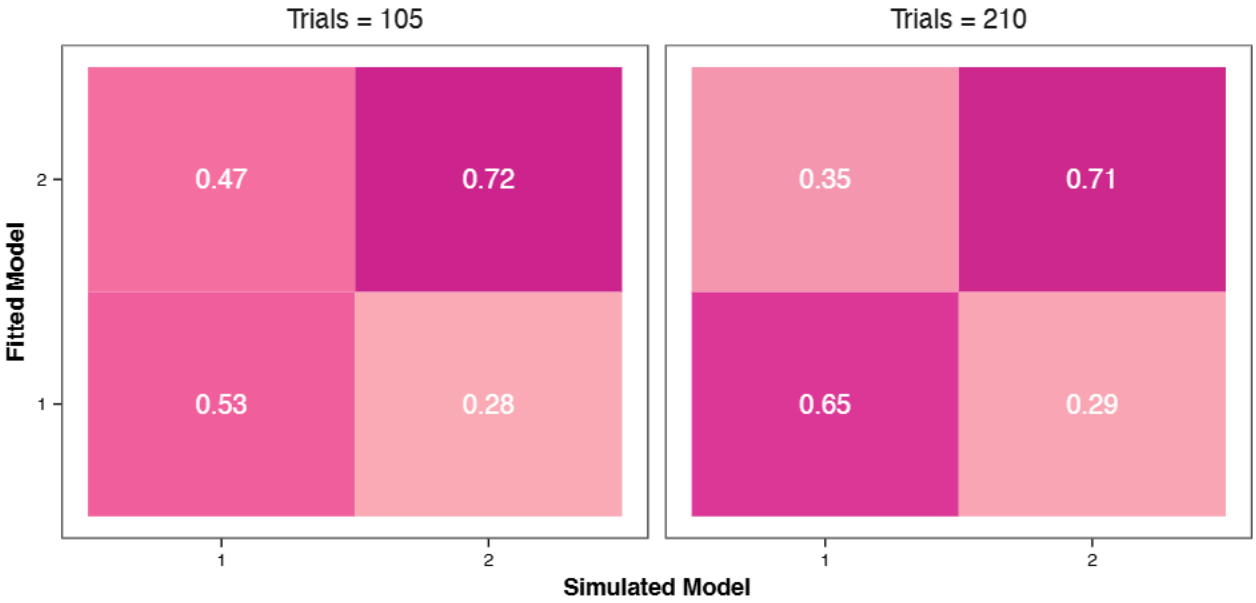


Figure S22. Model recovery. Percent of subjects better fit by the data generating model versus the alternative model using WAIC for datasets with 105 or 210 trials. Model 1: Differential Learning Model. Model 2: Differential Weight Model

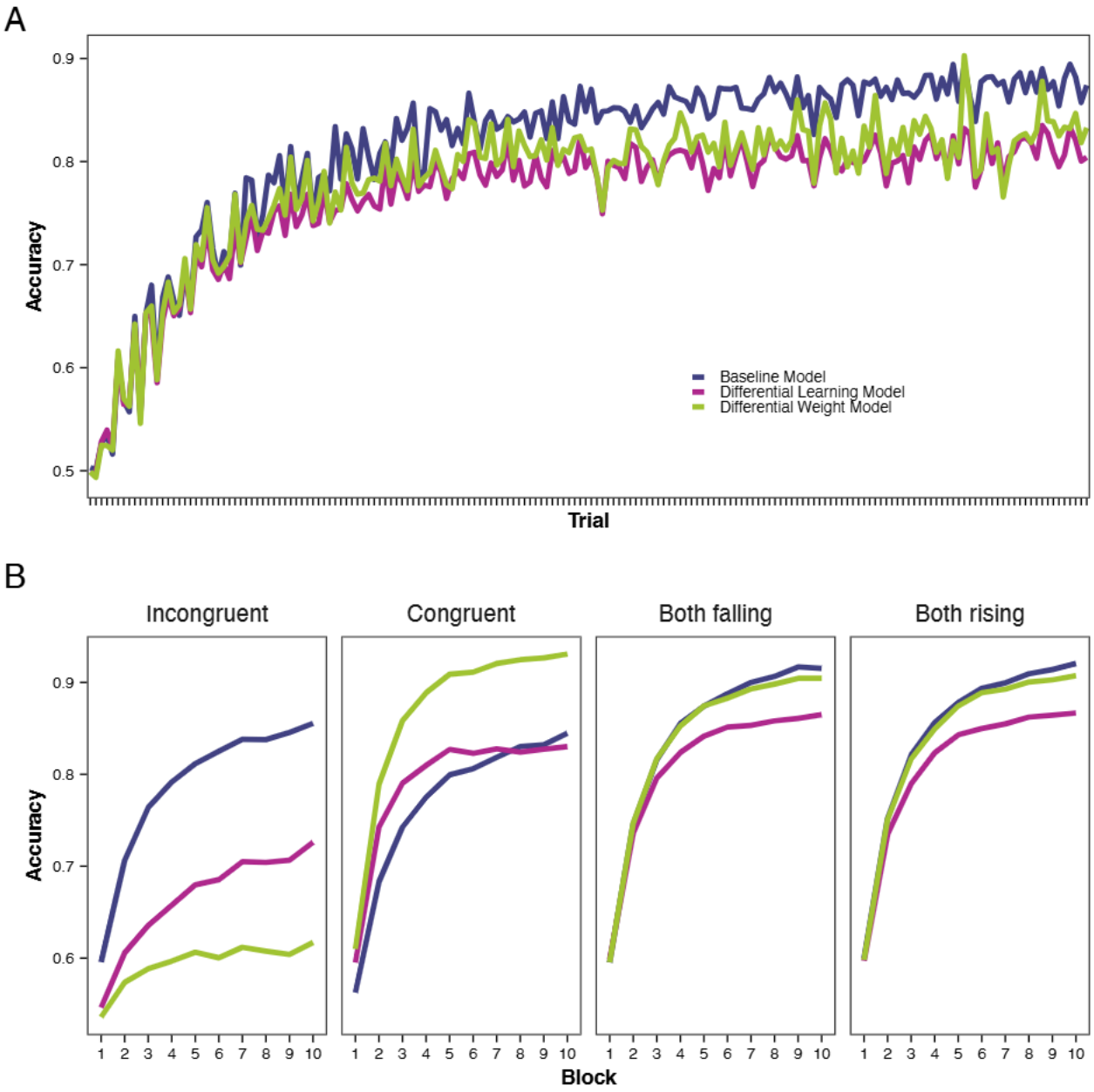


Figure S23. Model simulations with double the number of trials compared to the experiment. **(A,B)** Average choice accuracy in the model simulations using the mean posterior values across trials and subjects. **(A)** Experiment level. **(B)** Block level for each type of trial. Incongruent choice sets: worse option descending, better option ascending, Congruent choice sets: worse option ascending, better option descending. The model simulations is based 100 simulated datasets using the fitted parameter values for each subject in the Colors Task (Study 1) and Patterns Task (Study 2).

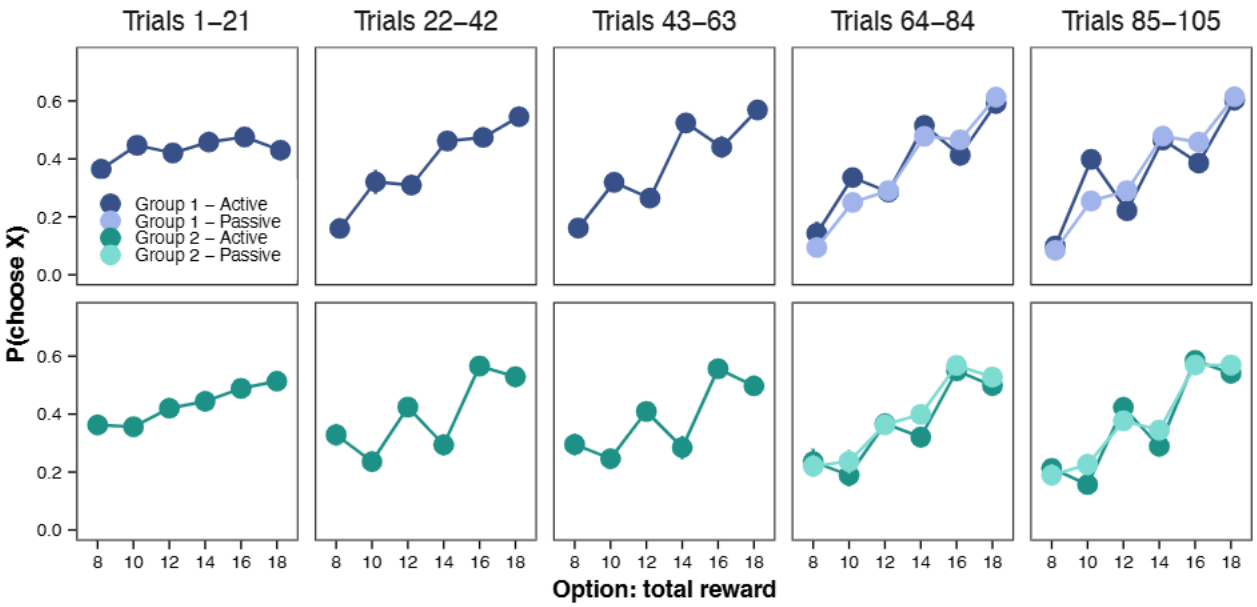


Figure S24. Behavioral bias over time in the passive learning condition. Probability of choosing a stimulus given it is in the choice set as a function of the total reward of the stimulus. In the first 63 trials, subjects did not make any decisions, but could learn passively from feedback shown on their screen. The feedback shown was based on the decisions of a matched partner from the main experiment. The feedback and timing of the feedback was the same as that of the matched partner, except that the subject could not see the foregone choice option on the choice screen. After trial 64, subjects made decisions. Active condition data is based on the matched subjects. Dots represent subject level averages and bars represent standard errors across subjects.

Supplementary Tables

	Choice (Left)				
	(patterns)	(colors)	(both)	(colors)	(passive)
	(1)	(2)	(3)	(4)	(5)
Right Desc	−0.49***	−1.35***	−0.50***	−1.75***	−0.53***
	(0.12)	(0.17)	(0.14)	(0.22)	(0.13)
Left Desc	0.30*	1.16***	0.31*	1.33***	0.86***
	(0.12)	(0.15)	(0.13)	(0.20)	(0.13)
Total (L-R)	1.73***	2.04***	1.72***	2.35***	1.67***
	(0.11)	(0.10)	(0.10)	(0.18)	(0.08)
Condition (Colors)			−0.06		
			(0.11)		
Right Desc:Condition			−0.84***		
			(0.20)		
Left Desc:Condition			0.84***		
			(0.19)		
Total (L-R):Condition			0.34*		
			(0.15)		
Passive				−0.17	
				(0.19)	
Right Desc:Passive				0.73*	
				(0.35)	
Left Desc:Passive				−0.23	
				(0.33)	
Total (L-R):Passive				0.33	
				(0.29)	
Constant	0.12	0.06	0.12	0.23	−0.12
	(0.08)	(0.08)	(0.08)	(0.13)	(0.12)
Observations	5,358	5,219	10,577	4,769	1,710
Log Likelihood	-2,631.18	-2,189.50	-4,827.39	-1,955.14	-807.41
Akaike Inf. Crit.	5,290.35	4,407.00	9,690.77	3,946.29	1,624.83
Bayesian Inf. Crit.	5,382.56	4,498.84	9,821.57	4,062.74	1,652.05
Note:					*p<0.05; **p<0.01; ***p<0.001

Table S1. Type of option effect on choice. Logistic regression of choice of left stimulus on whether the right and left stimuli are descending or ascending controlling for difference in mean experienced total reward between left and right stimuli. (1) Patterns Task. (2) Colors Task. (3) Patterns and Colors Tasks. (4) Colors Passive and Active Conditions. (5) Colors Passive Condition. Continuous variables are z-scored. Regressions include random intercepts and random slopes for whether the right and left stimuli are descending or ascending and for difference in mean experienced total reward at the subject level.

	Choice (Left)							
	(patterns)	(colors)	(both)	(patterns)	(colors)	(both)	(colors)	(passive)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Immediate (L-R)	1.54***	2.18***	1.53***	1.59***	2.24***	1.57***	2.56***	2.62***
	(0.11)	(0.11)	(0.10)	(0.11)	(0.11)	(0.10)	(0.18)	(0.26)
Delayed (L-R)	1.11***	0.90***	1.09***	1.14***	0.92***	1.12***	1.00***	1.48***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.14)	(0.21)
Trial				0.03	0.01	0.02		
				(0.04)	(0.04)	(0.04)		
Immediate (L-R):Trial				0.34***	0.38***	0.34***		
				(0.04)	(0.05)	(0.04)		
Delayed (L-R):Trial				0.19***	0.12**	0.19***		
				(0.04)	(0.04)	(0.04)		
Condition (Colors)			−0.06			−0.06		
			(0.07)			(0.07)		
Condition:Trial						−0.01		
						(0.05)		
Immediate (L-R):Condition			0.67***			0.68***		
			(0.15)			(0.15)		
Delayed (L-R):Condition			−0.17			−0.18		
			(0.14)			(0.14)		
Immediate (L-R):Condition:Trial						0.04		
						(0.07)		
Delayed (L-R):Condition:Trial						−0.07		
						(0.06)		
Passive							0.07	
							(0.11)	
Immediate (L-R):Passive							0.02	
							(0.29)	
Delayed (L-R):Passive							0.45	
							(0.23)	
Constant	0.02	−0.04	0.02	0.02	−0.03	0.02	0.01	0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)	(0.08)
Observations	5,358	5,219	10,577	5,358	5,219	10,577	4,769	1,710
Log Likelihood	-2,643.61	-2,195.38	-4,842.52	-2,606.39	-2,167.94	-4,777.88	-1,967.13	-682.78
Akaike Inf. Crit.	5,305.23	4,408.77	9,709.05	5,236.78	4,359.87	9,591.75	3,958.26	1,383.56
Bayesian Inf. Crit.	5,364.50	4,467.81	9,796.24	5,315.82	4,438.60	9,722.55	4,035.89	1,432.56

Note: *p<0.05; **p<0.01; ***p<0.001

Table S2. Type of reward effect on choice. Logistic regression of choice of left stimulus on difference in mean experienced immediate and delayed reward between left and right stimuli and interactions of these effects with trial number. (1,4) Patterns Task. (2,5) Colors Task. (3,6) Patterns and Colors Tasks. (7) Colors Task Passive and Active Learning Conditions. (8) Colors Task Passive Condition. Continuous variables are z-scored. Regressions include random intercepts and random slopes for difference in mean experienced immediate and delayed rewards between left and right stimuli at the subject level.

	log RT (s)		
	(patterns)	(colors)	(both)
	(1)	(2)	(3)
|VD Total| (L-R)	-0.04^{***}	-0.05^{***}	-0.04^{***}
	(0.01)	(0.01)	(0.01)
OV Total (L+R)	-0.05^{***}	-0.06^{***}	-0.05^{***}
	(0.01)	(0.01)	(0.01)
Condition (Colors)			0.66^{***}
			(0.03)
|VD Total| (L-R):Condition			-0.01
			(0.01)
OV Total (L+R):Condition			-0.01
			(0.01)
Constant	0.0004	0.66^{***}	0.0004
	(0.02)	(0.02)	(0.02)
Observations	5,358	5,219	10,577
Log Likelihood	-1,369.62	-2,018.76	-3,438.33
Akaike Inf. Crit.	2,759.24	4,057.53	6,902.66
Bayesian Inf. Crit.	2,825.10	4,123.13	6,997.13
Note:			*p<0.05; **p<0.01; ***p<0.001

Table S3. Value difference (VD) and overall value effects (OV) on response time. Linear regression of log RT on the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli. (1) Patterns Task. (2) Colors Task. (3) Patterns and Colors Tasks. Continuous variables are z-scored. Regressions include random intercepts and random slopes for the absolute difference in mean experiened reward and mean experienced total reward between left and right stimuli at the subject level.

	log RT (s)		
	(patterns)	(colors)	(both)
	(1)	(2)	(3)
VD Immediate (L-R)	−0.02***	−0.04***	−0.02***
	(0.01)	(0.01)	(0.01)
VD Delayed (L-R)	−0.01	−0.002	−0.01
	(0.01)	(0.01)	(0.01)
OV Immediate (L+R)	−0.04***	−0.07***	−0.04***
	(0.01)	(0.01)	(0.01)
OV Delayed (L+R)	−0.03***	−0.03***	−0.03***
	(0.01)	(0.01)	(0.01)
Condition (Colors)			0.66***
			(0.03)
VD Immediate (L-R):Condition			−0.02**
			(0.01)
VD Delayed (L-R):Condition			0.01
			(0.01)
OV Immediate (L+R):Condition			−0.03**
			(0.01)
OV Delayed (L+R):Condition			0.002
			(0.01)
Constant	0.0004	0.66***	0.0004
	(0.02)	(0.02)	(0.02)
Observations	5,358	5,219	10,577
Log Likelihood	-1,379.31	-1,988.68	-3,418.83
Akaike Inf. Crit.	2,800.63	4,019.36	6,889.66
Bayesian Inf. Crit.	2,938.94	4,157.12	7,078.58
Note:			*p<0.05; **p<0.01; ***p<0.001

Table S4. Value difference (VD) and overall value (OV) effects by type of reward on response time. Linear regression of log RT on the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli for both immediate and delayed rewards. (1) Patterns Task. (2) Colors Task. (3) Patterns and Colors Tasks. Continuous variables are z-scored. Regressions include random intercepts and random slopes for the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli for both immediate and delayed rewards at the subject level.

	log RT (s)		
	(patterns)	(colors)	(both)
	(1)	(2)	(3)
Condition (Colors)			0.68***
			(0.03)
Reward Total (L-R)	−0.002	−0.01***	−0.002
	(0.002)	(0.002)	(0.002)
Reward Total Squared (L-R)	−0.001***	−0.002***	−0.001***
	(0.0002)	(0.0002)	(0.0002)
Reward Total (L-R):Condition			−0.01**
			(0.003)
Reward Total Squared (L-R):Condition			−0.001*
			(0.0003)
Constant	0.04	0.72***	0.04
	(0.03)	(0.02)	(0.02)
Observations	3,211	3,131	6,342
Log Likelihood	-897.61	-1,291.13	-2,219.42
Akaike Inf. Crit.	1,809.22	2,596.26	4,458.84
Bayesian Inf. Crit.	1,851.74	2,638.61	4,526.39
Note:			*p<0.05; **p<0.01; ***p<0.001

Table S5. Value difference effects between descending and ascending options on response time. Linear regression of log RT on the absolute difference in underlying reward between descending and ascending stimuli and the square of the absolute difference in underlying reward between descending and ascending stimuli. (1) Patterns Task. (2) Colors Task. (3) Patterns and Colors Tasks. (1,2) Regressions include random intercepts and random slopes for the absolute difference in underlying reward between descending and ascending stimuli at the subject level.

	Dwell Proportion (I-D)				P(First Fix to I)			
	(patterns)	(patterns)	(inlab)	(inlab)	(patterns)	(patterns)	(inlab)	(inlab)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Immediate Points	−0.01	−0.0002	0.07***	0.08***	0.01	−0.02	0.06	0.12
	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)	(0.07)	(0.07)	(0.11)
Delayed Points	−0.02**	0.003	−0.09***	−0.09***	−0.04	0.15*	0.01	−0.10
	(0.01)	(0.02)	(0.01)	(0.01)	(0.04)	(0.08)	(0.07)	(0.11)
Immediate Stimulus Asc	−0.01	−0.001	0.20***	0.22***	0.06	−0.08	−0.12	0.20
	(0.01)	(0.03)	(0.03)	(0.04)	(0.07)	(0.14)	(0.20)	(0.31)
Delayed Stimulus Desc	0.01	0.02	−0.05***	−0.03	0.07	−0.27	−0.22**	−0.23
	(0.01)	(0.03)	(0.01)	(0.02)	(0.07)	(0.15)	(0.08)	(0.12)
Position (Immediate Bottom)				−0.05				−1.91***
				(0.05)				(0.35)
Immediate Points:Position				−0.02				−0.10
				(0.02)				(0.14)
Delayed Points:Position				0.005				0.19
				(0.02)				(0.15)
Immediate Stimulus Asc:Position				−0.04				−0.56
				(0.06)				(0.40)
Delayed Stimulus Asc:Position				−0.03				−0.001
				(0.03)				(0.15)
Constant	0.05	−0.02	0.002	0.02	0.09	0.31	0.98***	1.85***
	(0.03)	(0.05)	(0.03)	(0.04)	(0.11)	(0.23)	(0.24)	(0.25)
Observations	7,795	1,863	4,763	4,763	7,720	1,845	4,717	4,717
Log Likelihood	-4,442.95	-1,290.41	-2,687.06	-2,697.90	-4,914.60	-1,170.21	-2,372.51	-2,352.47
Akaike Inf. Crit.	8,899.90	2,594.82	5,388.12	5,419.80	9,841.19	2,352.42	4,757.02	4,726.94
Bayesian Inf. Crit.	8,948.63	2,633.53	5,433.40	5,497.43	9,882.90	2,385.54	4,795.77	4,797.99

Note: *p<0.05; **p<0.01; ***p<0.001

Table S6. Attention and reward feedback. (1,2,3,4) Linear regression of difference in proportion of dwell time within a trial between immediate and delayed reward as a function of the size and type (descending vs. ascending) of reward. (1) Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. (2) Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. (3,4) In-lab Eye-tracking Study. (5,6,7,8) Logistic regression of first fixation location (immediate vs. delayed) as a function of the size and type (descending vs. ascending) of reward. (5) Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. (6) Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. (7,8) In-lab Eye-tracking Study. Continuous variables are z-scores. Regressions include random intercept at the subject level.

	Dwell Proportion (I-D)			P(First Fix to I)		
	(patterns)	(patterns)	(inlab)	(patterns)	(patterns)	(inlab)
	(1)	(2)	(3)	(4)	(5)	(6)
|Prediction Error Immediate|	−0.001	0.01*	−0.01**	0.02	0.04	−0.01
	(0.002)	(0.01)	(0.004)	(0.01)	(0.03)	(0.02)
|Prediction Error Delayed|	−0.003	0.01	−0.005	0.004	−0.01	−0.01
	(0.002)	(0.005)	(0.002)	(0.01)	(0.02)	(0.02)
Predicted Value Immediate	−0.001	−0.01	0.01*	−0.01	−0.02	0.05**
	(0.002)	(0.005)	(0.003)	(0.01)	(0.02)	(0.02)
Predicted Value Delayed	−0.01**	−0.002	0.02***	−0.01	0.05	0.02
	(0.002)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)
Constant	0.09**	−0.03	0.03	0.21	0.02	0.67*
	(0.03)	(0.07)	(0.04)	(0.14)	(0.31)	(0.27)
Observations	7,634	1,845	4,716	7,634	1,845	4,716
Log Likelihood	-4,320.62	-1,274.15	-2,582.94	-4,846.06	-1,169.72	-2,372.63
Akaike Inf. Crit.	8,657.24	2,564.30	5,179.88	9,706.12	2,353.44	4,757.26
Bayesian Inf. Crit.	8,712.76	2,608.46	5,225.09	9,754.70	2,392.08	4,796.01
Note:						*p<0.05; **p<0.01; ***p<0.001

Table S7. Attention and model predictions. (1,2,3) Linear regression of difference in proportion of dwell time within a trial between immediate and delayed reward as a function of the absolute prediction error and the predicted values for the immediate and delayed rewards derived from the differential learning model. (1) Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. (2) Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. (3) In-lab Eye-tracking Study. (4,5,6) Logistic regression of first fixation location (immediate vs. delayed) as a function of the absolute prediction error and the predicted values for the immediate and delayed rewards derived from the differential learning model. Continuous variables are z-scores. Regressions include random intercept at the subject and trial level. (4) Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. (5) Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top. (6) In-lab Eye-tracking Study.

	Choice (Left)	
	(patterns)	(patterns)
	(1)	(2)
Immediate (L-R)	1.25***	1.31***
	(0.09)	(0.16)
Delayed (L-R)	0.72***	0.48***
	(0.08)	(0.11)
Dwell Proportion (L-R)	0.43***	0.49***
	(0.07)	(0.11)
Constant	−0.08	0.12
	(0.06)	(0.10)
Observations	7,698	1,831
Log Likelihood	-4,276.45	-1,017.00
Akaike Inf. Crit.	8,580.91	2,054.00
Bayesian Inf. Crit.	8,678.19	2,109.13
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table S8. Dwell proportion difference on choice. (1,2) Logistic regression of choice of left stimulus on difference in mean experienced reward between left and right stimulus for immediate and delayed rewards and the proportion of dwell time between left and right options within a trial. Continuous variables are z-scored. Regressions include random intercept and slopes for difference in mean experienced reward between left and right stimulus for immediate and delayed rewards and the proportion of dwell time between left and right options within a trial at the subject level. (1) Feedback Position Condition: Immediate Reward: Top - Delayed Reward: Bottom. (2) Feedback Position Condition: Immediate Reward: Bottom - Delayed Reward: Top.

	Mean Error		Incongruent Set Error	
	(patterns)	(inlab)	(patterns)	(inlab)
	(1)	(2)	(3)	(4)
Survey 1 Rankings Error	0.04***	0.03***	0.04**	0.03
	(0.01)	(0.01)	(0.01)	(0.02)
Survey 2 Reward Comparison Error	0.01	0.001	−0.07**	−0.11**
	(0.01)	(0.01)	(0.03)	(0.03)
Survey 3 Points Error	0.003		0.01	
	(0.01)		(0.02)	
Survey 3 Points Error Immediate		0.05		0.01
		(0.03)		(0.07)
Survey 3 Points Error Delayed		0.04		0.15*
		(0.02)		(0.06)
Constant	0.13**	−0.04	0.41***	0.18
	(0.04)	(0.08)	(0.09)	(0.20)
Observations	73	46	73	46
R^2	0.40	0.54	0.19	0.39
Adjusted R^2	0.37	0.50	0.15	0.33
Residual Std. Error	0.09 (df = 69)	0.08 (df = 41)	0.20 (df = 69)	0.21 (df = 41)
F Statistic	15.21*** (df = 3; 69)	12.22*** (df = 4; 41)	5.40** (df = 3; 69)	6.62*** (df = 4; 41)
Note:				*p<0.05; **p<0.01; ***p<0.001

Table S9. Choice and memory errors. (1,2) Linear regression of mean errors at the subject level on the memory error for each survey. (1) Colors Task. (2) In-lab Eye-tracking. (3,4) Linear regression of mean errors for incongruent choice sets (worse option descending, better option ascending) at the subject level on the memory error for each survey. (3) Colors Task. (4) In-lab Eye-tracking.

	Bias Choice		Bias Learning	
	(patterns)	(inlab)	(patterns)	(inlab)
	(1)	(2)	(3)	(4)
Survey 1 Rankings Error	0.05	−0.19*	−0.004	−0.003
	(0.06)	(0.08)	(0.01)	(0.01)
Survey 2 Reward Comparison Error	−0.44***	−0.28*	−0.04	−0.05*
	(0.12)	(0.14)	(0.02)	(0.02)
Survey 3 Points Error	−0.01		−0.005	
	(0.07)		(0.02)	
Survey 3 Points Error Immediate		−0.61*		−0.01
		(0.27)		(0.05)
Survey 3 Points Error Delayed		0.08		0.03
		(0.25)		(0.04)
Constant	1.46***	3.89***	0.20*	0.19
	(0.40)	(0.78)	(0.08)	(0.13)
Observations	73	46	73	46
R^2	0.17	0.35	0.05	0.15
Adjusted R^2	0.14	0.29	0.01	0.06
Residual Std. Error	0.92 (df = 69)	0.84 (df = 41)	0.19 (df = 69)	0.14 (df = 41)
F Statistic	4.82** (df = 3; 69)	5.62** (df = 4; 41)	1.32 (df = 3; 69)	1.76 (df = 4; 41)
Note:				*p<0.05; **p<0.01; ***p<0.001

Table S10. Bias and memory errors. (1) Linear regression of behavioral bias at the subject level on the memory error for each survey. The behavioral bias is the difference in coefficients of mean experienced immediate and delayed reward from the mixed effects logistic regression on choice. (2) Linear regression of difference in mean posterior of the learning rates between immediate and delayed rewards on the memory error for each survey.

	Bias Choice					Bias Learning				
	(colors)	(colors)	(colors)	(colors)	(inlab)	(colors)	(colors)	(colors)	(colors)	(inlab)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Accuracy 2-back	−2.85					−0.55*				
	(1.57)					(0.23)				
D' 2-back			−0.31*					−0.05*		
			(0.14)					(0.02)		
Accuracy 3-back		−0.89					−0.07			
		(1.48)					(0.25)			
D' 3-back				−0.12					−0.02	
				(0.21)					(0.03)	
Accuracy visual					0.65					0.32
					(1.28)					(0.18)
Constant	3.87**	1.70	2.14***	1.19***	1.35	0.64**	0.17	0.29***	0.15**	−0.08
	(1.46)	(1.12)	(0.43)	(0.32)	(0.79)	(0.22)	(0.19)	(0.06)	(0.05)	(0.11)
Observations	86	51	86	51	46	86	51	86	51	46
R^2	0.04	0.01	0.06	0.01	0.01	0.06	0.002	0.07	0.01	0.07
Adjusted R^2	0.03	-0.01	0.04	-0.01	-0.02	0.05	-0.02	0.06	-0.01	0.05
Note:										*p<0.05; **p<0.01; ***p<0.001

Table S11. Bias and n-back working memory task performance. Linear regression of bias at the subject level on performance on the 2-back task, 3-back task and the change localization task. (1,2,5) Behavioral bias on accuracy. (3,4) Behavioral bias on d prime. (6,7,10) Difference in mean posterior of the learning rates between immediate and delayed rewards on accuracy. (8,9) Difference in mean posterior of the learning rates between immediate and delayed rewards on d prime. (1,2,6,7) Accuracy is measured as the correct responses (both correctly identified targets as well as correctly identified non-targets) out of total number of stimuli. (5,10) Accuracy is measured as the percent correct responses in the change localization task. (3,4,8,9) D prime is the sensitivity measure calculated as the difference in z-scored hit rates and z-scored false alarm rates. Hit rate refers to the proportion of correctly identified target stimuli (i.e., stimuli that match the one presented n items back). It is calculated as the number of correct responses divided by the total number of target stimuli. False alarm rate refers to the proportion of non-target stimuli that are incorrectly identified as targets. It is calculated as the number of false alarms divided by the total number of non-target stimuli. These measures where corrected to account for cases in which the hit rate equals 1 or false alarm rate equals 0.

	Today-1 Month		Today-6 Months		1 Month-6 Months	
	(colors)	(colors)	(colors)	(colors)	(colors)	(colors)
	(1)	(2)	(3)	(4)	(5)	(6)
Bias Choice	2.04*		1.86		1.53	
	(0.95)		(1.10)		(1.04)	
Bias Learning		11.04		8.03		3.63
		(6.41)		(7.45)		(7.04)
Constant	109.69***	110.77***	116.74***	117.99***	115.81***	117.22***
	(1.59)	(1.37)	(1.85)	(1.60)	(1.75)	(1.51)
Observations	87	87	87	87	87	87
R^2	0.05	0.03	0.03	0.01	0.02	0.003
Adjusted R^2	0.04	0.02	0.02	0.002	0.01	-0.01
Residual Std. Error (df = 85)	10.15	10.24	11.79	11.91	11.13	11.25
F Statistic (df = 1; 85)	4.62*	2.97	2.84	1.16	2.15	0.27
Note:						*p<0.05; **p<0.01; ***p<0.001

Table S12. Bias and intertemporal preferences. (1, 3, 5) Linear regression of intertemporal choice indifference point on behavioral bias at the subject level. The behavioral bias is the difference in coefficients of mean experienced immediate and delayed rewards from the mixed effects logistic regression on choice. (2, 4, 6) Linear regression of intertemporal choice indifference point on difference in mean posterior of the learning rates between immediate and delayed rewards.

	Today-1 Month		Today-6 Months		1 Month-6 Months	
	(inlab)	(inlab)	(inlab)	(inlab)	(inlab)	(inlab)
	(1)	(2)	(3)	(4)	(5)	(6)
Bias Choice	−2.00		−1.94		−1.48	
	(1.61)		(1.79)		(1.74)	
Bias Learning		−10.34		−11.13		−19.45
		(11.29)		(12.50)		(11.88)
Constant	116.22***	113.87***	122.43***	120.26***	120.44***	120.00***
	(3.22)	(2.02)	(3.57)	(2.23)	(3.49)	(2.12)
Observations	46	46	46	46	46	46
R^2	0.03	0.02	0.03	0.02	0.02	0.06
Adjusted R^2	0.01	-0.004	0.004	-0.005	-0.01	0.04
Residual Std. Error (df = 44)	10.71	10.80	11.90	11.95	11.61	11.37
F Statistic (df = 1; 44)	1.55	0.84	1.18	0.79	0.72	2.68
Note:						*p<0.05; **p<0.01; ***p<0.001

Table S13. Bias and intertemporal preferences in-lab eye-tracking study. (1, 3, 5) Linear regression of intertemporal choice indifference point on behavioral bias at the subject level. The behavioral bias is the difference in coefficients of mean experienced immediate and delayed rewards from the mixed effects logistic regression on choice. (2, 4, 6) Linear regression of intertemporal choice indifference point on difference in mean posterior of the learning rates between immediate and delayed rewards.

	Choice (Left)					
	(combined)	(separate)	(combined)	(combined)	(separate)	(separate)
	(1)	(2)	(3)	(4)	(5)	(6)
Right Desc	−0.45***	−0.39***	−0.67***	1.51***	−0.57*	1.74***
	(0.07)	(0.10)	(0.17)	(0.26)	(0.24)	(0.34)
Left Desc	0.45***	0.29**	0.56***	−1.55***	0.37	−1.80***
	(0.07)	(0.10)	(0.17)	(0.26)	(0.22)	(0.34)
Total (L-R)	1.60***	1.58***		2.22***		2.34***
	(0.12)	(0.17)		(0.22)		(0.30)
Position (Immediate Bottom)		−0.01			−0.17	−0.21
		(0.15)			(0.18)	(0.24)
Right Desc:Position		−0.12			−0.21	−0.52
		(0.15)			(0.35)	(0.50)
Left Desc:Position		0.32*			0.39	0.57
		(0.15)			(0.33)	(0.50)
Total (L-R):Position		0.06				−0.28
		(0.23)				(0.43)
Constant	−0.01	−0.001	0.12	0.09	0.20	0.18
	(0.08)	(0.10)	(0.09)	(0.12)	(0.12)	(0.17)
Observations	4,740	4,740	2,799	2,799	2,799	2,799
Log Likelihood	-2,433.78	-2,430.14	-1,787.52	-1,496.03	-1,786.22	-1,494.68
Akaike Inf. Crit.	4,881.55	4,882.29	3,593.03	3,020.06	3,596.45	3,025.36
Bayesian Inf. Crit.	4,926.80	4,953.39	3,646.46	3,103.18	3,667.69	3,132.23
Note:						*p<0.05; **p<0.01; ***p<0.001

Table S14. Type of option effect on choice by feedback position. Logistic regression of choice of left stimulus on whether the right and left stimuli are descending or ascending controlling for difference in mean experienced total reward between left and right stimuli. (1,3,4) Combined position conditions. (2,5,6) Separate position conditions. (1,2) Online Patterns Study. (3,4,5,6) In-lab Eye-tracking Study. Continuous variables are z-scored. Regressions include random intercepts and random slopes for whether the right and left stimuli are descending or ascending and for difference in mean experienced total reward at the subject level.

	Error			
	(combined)	(separate)	(combined)	(separate)
	(1)	(2)	(3)	(4)
Set Type Congruent	-0.17^{***}	-0.14^{**}	-0.25^{***}	-0.19^{**}
	(0.03)	(0.05)	(0.04)	(0.06)
Position (Immediate Bottom)		0.01		0.10
		(0.05)		(0.06)
Set Type Congruent:Position		-0.07		-0.13
		(0.06)		(0.09)
Constant	0.40^{***}	0.39^{***}	0.43^{***}	0.39^{***}
	(0.02)	(0.03)	(0.03)	(0.04)
Observations	160	160	94	94
R^2	0.15	0.16	0.26	0.28
Adjusted R^2	0.15	0.14	0.25	0.26
Residual Std. Error	0.20 (df = 158)	0.20 (df = 156)	0.21 (df = 92)	0.21 (df = 90)
F Statistic	28.31^{***} (df = 1; 158)	9.88^{***} (df = 3; 156)	31.77^{***} (df = 1; 92)	11.63^{***} (df = 3; 90)
Note:				*p<0.05; **p<0.01; ***p<0.001

Table S15. Type of choice set effect on error rate by feedback position. Linear regression of type of choice set (incongruent or congruent) on mean error rate. (1,3) Combined position conditions. (2,4) Separate position conditions. (1,2) Online Patterns Study. (3,4) In-lab Eye-tracking Task.

	Choice (Left)							
	(combined)	(separate)	(combined)	(separate)	(combined)	(separate)	(combined)	(separate)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Immediate (L-R)	1.60***	1.48***	1.52***	1.53***	1.49***	1.42***	1.53***	1.47***
	(0.24)	(0.15)	(0.11)	(0.16)	(0.15)	(0.20)	(0.15)	(0.20)
Delayed (L-R)	0.72***	1.03***	1.03***	1.06***	−0.35***	−0.30*	−0.46***	−0.41**
	(0.16)	(0.15)	(0.11)	(0.15)	(0.10)	(0.14)	(0.11)	(0.14)
Trial			0.08*	0.08			−0.02	0.07
			(0.04)	(0.05)			(0.05)	(0.07)
Immediate (L-R):Trial			0.34***	0.40***			0.29***	0.30***
			(0.04)	(0.06)			(0.06)	(0.08)
Delayed (L-R):Trial			0.06	0.08			−0.25***	−0.23**
			(0.04)	(0.06)			(0.05)	(0.07)
Position (Immediate Bottom)		0.08		0.08		−0.22		−0.24
		(0.10)		(0.10)		(0.17)		(0.17)
Position:Trial				0.003				−0.19
				(0.08)				(0.10)
Immediate (L-R):Position		0.004		−0.02		0.15		0.14
		(0.22)		(0.22)		(0.29)		(0.30)
Delayed (L-R):Position		−0.04		−0.05		−0.11		−0.12
		(0.21)		(0.21)		(0.20)		(0.21)
Immediate (L-R):Position:Trial				−0.11				−0.002
				(0.09)				(0.12)
Delayed (L-R):Position:Trial				−0.03				−0.04
				(0.08)				(0.11)
Constant	0.19*	−0.03	0.02	−0.02	0.09	0.20	0.09	0.20
	(0.09)	(0.07)	(0.05)	(0.07)	(0.09)	(0.11)	(0.09)	(0.11)
Observations	1,077	4,740	4,740	4,740	2,799	2,799	2,799	2,799
Log Likelihood	-532.01	-2,370.70	-2,337.84	-2,336.77	-1,479.77	-1,478.60	-1,461.08	-1,458.07
Akaike Inf. Crit.	1,082.02	4,765.40	4,699.68	4,709.54	2,977.53	2,981.20	2,946.15	2,952.14
Bayesian Inf. Crit.	1,126.86	4,842.96	4,777.25	4,825.89	3,030.96	3,052.45	3,017.39	3,059.00

Note: *p<0.05; **p<0.01; ***p<0.001

Table S16. Type of reward effect on choice by feedback position. Logistic regression of choice of left stimulus on difference in mean experienced immediate and delayed rewards between left and right stimuli and interactions of these effects with trial number. (1,3,5,7) Combined position conditions. (2,4,6,8) Separate position conditions. (1,2,3,4) Online Patterns Study. (5,6,7,8) In-lab Eye-tracking Study. Continuous variables are z-scored. Regressions include random intercepts and random slopes for difference in mean experienced immediate and delayed rewards between left and right stimuli at the subject level.

	log RT (s)			
	(combined)	(separate)	(combined)	(separate)
	(1)	(2)	(3)	(4)
VD Total (L-R)	-0.03^{***}	-0.03^{***}	-0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
OV Total (L+R)	-0.04^{***}	-0.04^{***}	-0.03^{**}	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Position (Immediate Bottom)		0.06		-0.01
		(0.05)		(0.07)
VD Total (L-R):Position		0.005		-0.02
		(0.01)		(0.02)
OV Total (L+R):Position		0.001		-0.04^{*}
		(0.01)		(0.02)
Constant	0.01	-0.02	0.14^{***}	0.14^{**}
	(0.03)	(0.04)	(0.04)	(0.05)
Observations	4,740	4,740	2,799	2,799
Log Likelihood	-1,521.34	-1,529.24	-741.15	-746.76
Akaike Inf. Crit.	3,062.68	3,084.49	1,502.30	1,519.52
Bayesian Inf. Crit.	3,127.32	3,168.52	1,561.67	1,596.70
Note:				*p<0.05; **p<0.01; ***p<0.001

Table S17. Value difference (VD) and overall value effects (OV) on response time by feedback position. Linear regression of log RT on the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli. (1,3) Combined position conditions. (2,4) Separate position conditions. (1,2) Online Patters Study. (3,4) In-lab Eye-tracking Study. Continuous variables are z-scored. Regressions include random intercepts and random slopes for the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli at the subject level.

	log RT (s)			
	(combined)	(separate)	(combined)	(separate)
	(1)	(2)	(3)	(4)
VD Immediate (L-R)	-0.03^{***}	-0.03^{***}	-0.04^{***}	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
VD Delayed (L-R)	-0.02^{**}	-0.02^{*}	0.01^{*}	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
OV Immediate (L+R)	-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.01
	(0.005)	(0.01)	(0.01)	(0.01)
OV Delayed (L+R)	-0.03^{***}	-0.03^{***}	0.02^{**}	0.01
	(0.005)	(0.01)	(0.01)	(0.01)
Position (Immediate Bottom)		0.06		-0.01
		(0.05)		(0.07)
VD Immediate (L-R):Position		0.002		-0.05^{***}
		(0.01)		(0.01)
VD Delayed (L-R):Position		0.001		0.01
		(0.01)		(0.01)
OV Immediate (L+R):Position		-0.01		-0.04^{***}
		(0.01)		(0.01)
OV Delayed (L+R):Position		0.01		0.03^{*}
		(0.01)		(0.01)
Constant	0.01	-0.02	0.14^{***}	0.14^{**}
	(0.03)	(0.04)	(0.04)	(0.05)
Observations	4,740	4,740	2,799	2,799
Log Likelihood	-1,571.04	-1,586.22	-771.04	-769.58
Akaike Inf. Crit.	3,156.08	3,196.44	1,556.08	1,563.17
Bayesian Inf. Crit.	3,201.33	3,274.01	1,597.64	1,634.41
Note:				*p<0.05; **p<0.01; ***p<0.001

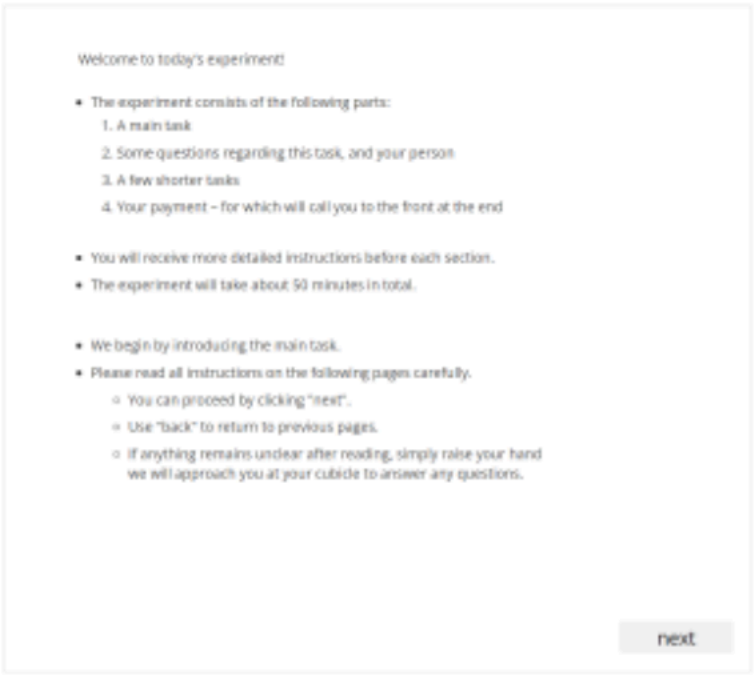
Table S18. Value difference (VD) and overall value (OV) effects by type of reward on response time by feedback position. Linear regression of log RT on the absolute difference in mean experienced reward and mean experienced total reward between left and right stimuli for both immediate and delayed rewards. (1,3) Combined position conditions. (2,4) Separate position conditions. (1,2) Online Patterns Study. (3,4) In-lab Eye-tracking Study. Continuous variables are z-scored. Regressions include random intercepts at the subject level.

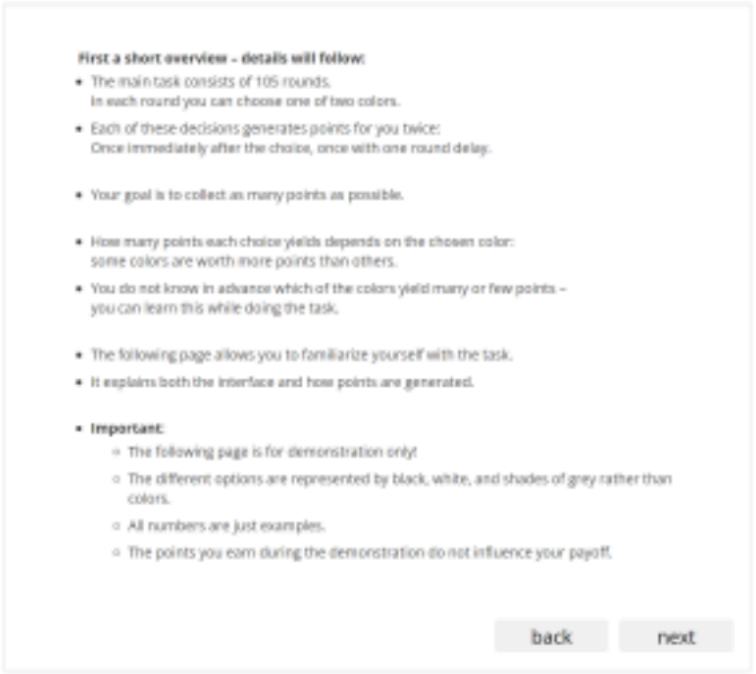
	log RT (s)			
	(combined)	(separate)	(combined)	(separate)
	(1)	(2)	(3)	(4)
Position (Immediate Bottom)		0.06		0.001
		(0.06)		(0.07)
Total (L-R)	−0.002	−0.002	−0.01*	−0.0004
	(0.002)	(0.003)	(0.003)	(0.004)
Total Squared (L-R)	−0.001***	−0.001**	−0.001***	−0.002***
	(0.0002)	(0.0003)	(0.0003)	(0.0004)
Total (L-R):Position		−0.001		−0.01**
		(0.004)		(0.01)
Total Squared (L-R):Position		−0.0001		0.0002
		(0.0004)		(0.001)
Constant	0.03	−0.005	0.18***	0.18***
	(0.03)	(0.04)	(0.04)	(0.05)
Observations	2,851	2,851	1,679	1,679
Log Likelihood	-998.66	-1,011.42	-370.60	-379.73
Akaike Inf. Crit.	2,011.32	2,042.83	755.21	779.45
Bayesian Inf. Crit.	2,053.00	2,102.39	793.19	833.71
Note:				*p<0.05; **p<0.01; ***p<0.001

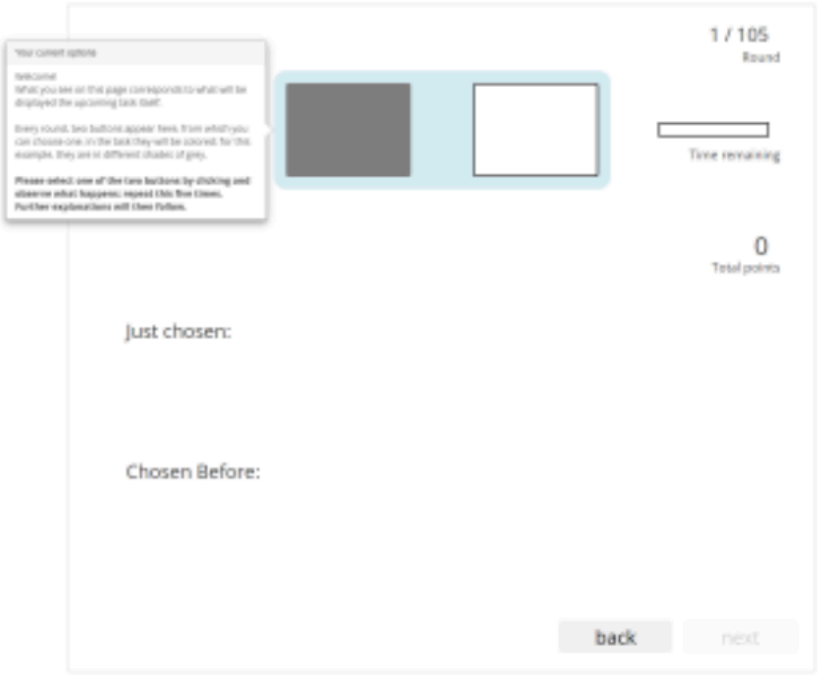
Table S19. Value difference effects between descending and ascending options on response time by feedback position. Linear regression of log RT on the absolute difference in underlying reward between descending and ascending stimuli and the square of the absolute difference in underlying reward between descending and ascending stimuli. (1,3) Combined position conditions. (2,4) Separate position conditions. (1,2) Online Patterns Task. (3,4) In-lab Eye-tracking Task. Regressions include random intercepts and random slope for the absolute difference in underlying reward between descending and ascending stimuli at the subject level.

Supplementary Materials

Instructions: Colors Task

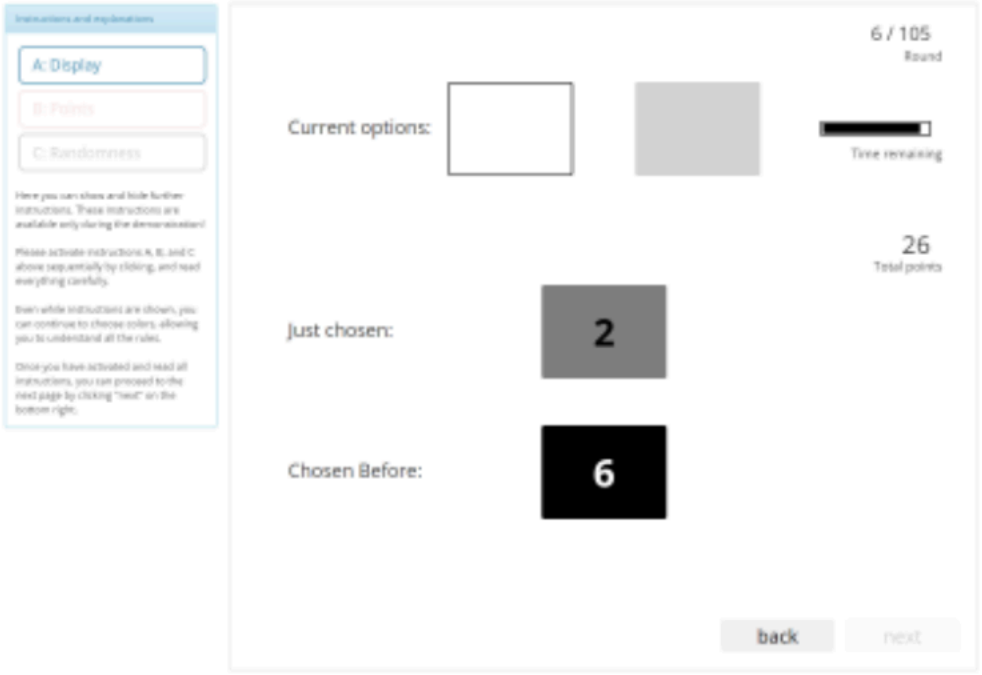


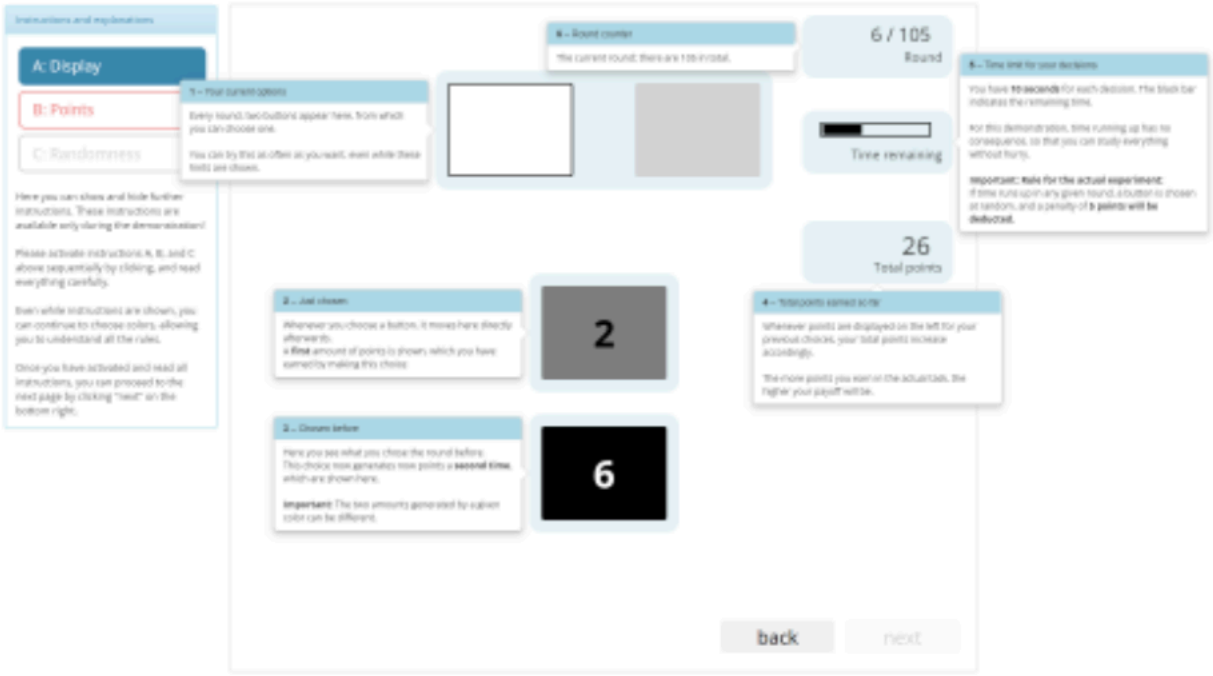


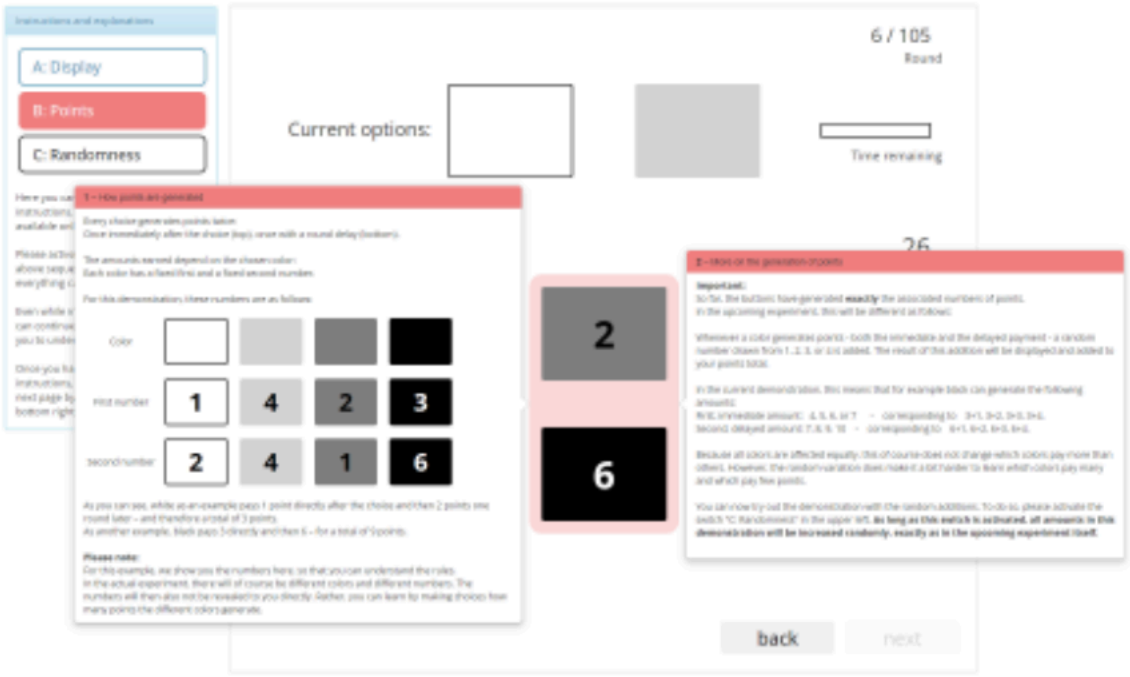

Your current options
Welcome
What you see on this page corresponds to what will be displayed the upcoming task itself.
Every round, two buttons appear here, from which you can choose one. In the task they will be colored, for this example, they are in different shades of grey.
Please select one of the two buttons by clicking and observe what happens; repeat this five times. Further explanations will then follow.
1 / 105
Round
Time remaining
0
Total points
Just chosen:
Chosen Before:
back
next

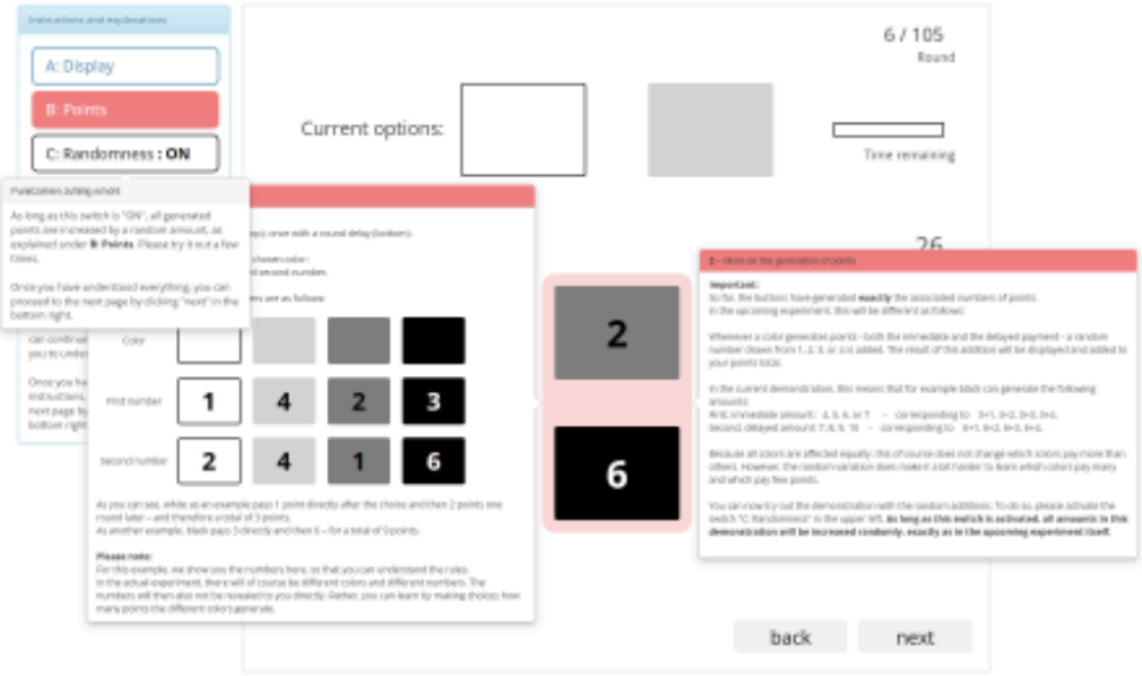

Your current options
Welcome
What you see on this page corresponds to what will be displayed the upcoming task itself.
Every round, two buttons appear here, from which you can choose one. In the task they will be colored, for this example, they are in different shades of grey.
Please select one of the two buttons by clicking and observe what happens; repeat this five times. Further explanations will then follow.
1 / 105
Round
Time remaining
0
Total points
Just chosen:
Chosen Before:
Please read and follow all of the instructions before proceeding.
back
next

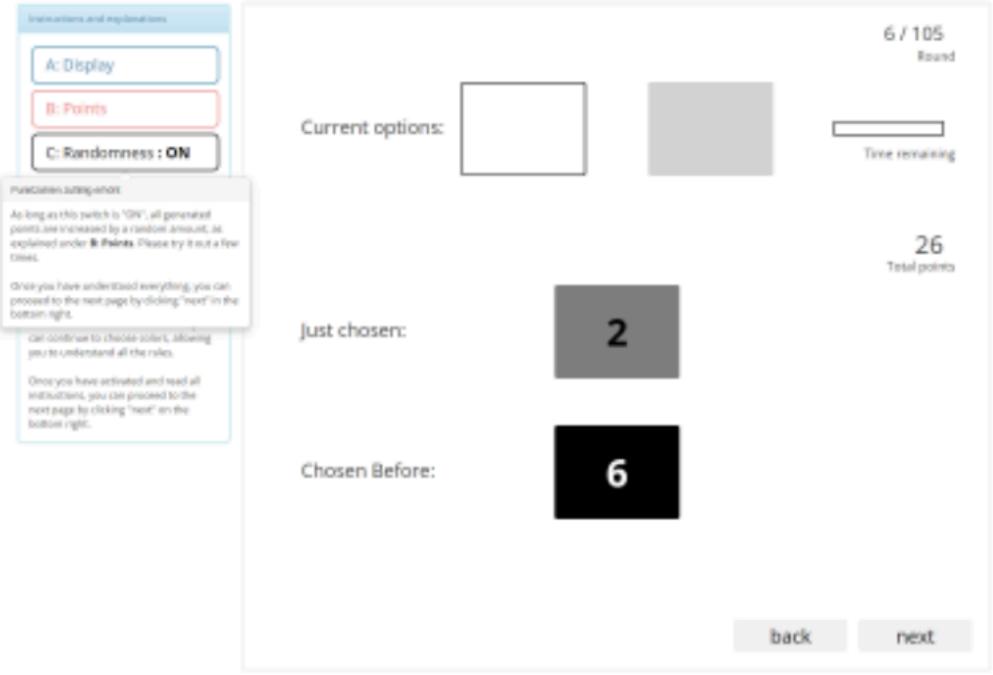

Your current options
Welcome!
What you see on this page corresponds to what will be displayed the upcoming task itself.
Every round, two buttons appear here, from which you can choose one. In the task they will be colored, for this example, they are in different shades of grey.
Please select one of the two buttons by clicking and observe what happens; repeat this five times.
Further explanations will then follow.
4 / 105
Round
Time remaining
9
Total points
Just chosen:
3
Chosen Before:
2
back
next


Instructions and explanations
A: Display
B: Points
C: Randomness
Here you can show and hide further instructions. These instructions are available only during the demonstration!
Please activate instructions A, B, and C above sequentially by clicking, and read everything carefully.
Even while instructions are shown, you can continue to choose colors, allowing you to understand all the rules.
Once you have activated and read all instructions, you can proceed to the next page by clicking "next" in the bottom right.
6 / 105
Round
Current options:
Time remaining
26
Total points
Just chosen:
2
Chosen Before:
6
back
next


Instructions and explanations
A: Display
B: Points
C: Randomness
6 / 105
Round
Time remaining
26
Total points
2
6
back
next


Instructions and explanations
A: Display
B: Points
C: Randomness
6 / 105
Round
Current options:
Time remaining
Color
First number
Second number
1
4
2
3
2
4
1
6
2
6
back
next


A: Display
B: Points
C: Randomness : ON
6 / 105
Round
Current options:
Time remaining
2
6
back
next


A: Display
B: Points
C: Randomness : ON
6 / 105
Round
Current options:
Time remaining
26
Total points
Just chosen:
2
Chosen Before:
6
back
next

To summarize:

- In each round, you choose one of two buttons.
- Each choice generates points twice.
- The amounts generated depend on the chosen color.
  First and second amounts generated by each color can be different.
- On top, 1, 2, 3 oder 4 points are always added at random – independently of the color.
- Your goal is to earn as many points as possible throughout the 105 rounds.
- To this end, you can learn bit by bit which colors yield more points than others.
- In each round, you have only 10 seconds to observe the latest numbers and choose the next color.
-
- A bar indicates the remaining time.
- If you do not choose in time, a color is picked at random for you.
  In addition, you incur a penalty of 5 points.

back next

- These are the six colors that are used in the experiment:

- If you have trouble differentiating any two of these colors:
  Please raise your hand **now** - an experimenter will come to your desk!
- Which two buttons are available in any given round does not depend on your past choices.
- It can happen that both buttons have the same color – this is on purpose.
  In this case it makes no difference which of the two you click.

back next

- Your payoff for this experiment depends considerably on how many points you earn in the upcoming task.
  - Thus, please make sure you understand all rules well before clicking on "start" below.
  - Before you start, you can go back in the instructions and also re-play the example, if you like.
  - If anything remains unclear, please let us know – we will come to your desk.

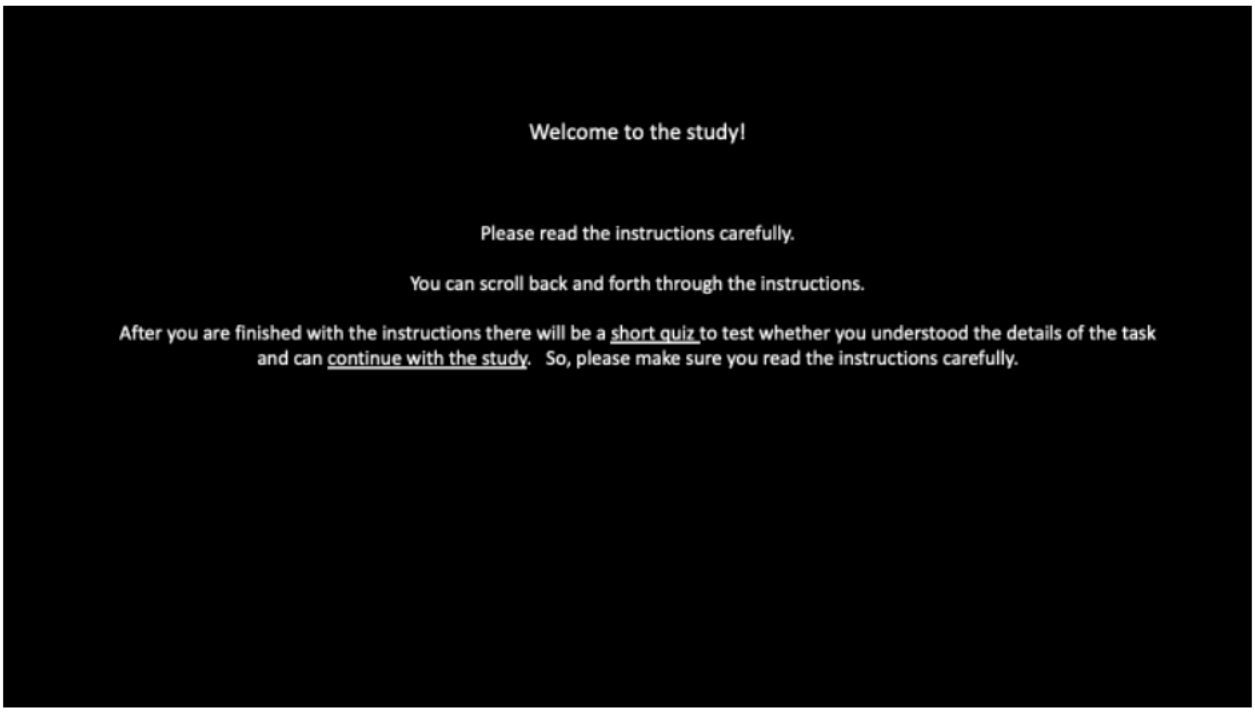
Your payment for this part of the experiment is calculated as follows:

- You receive **5 cents** for each point you earn **above 1800 points**.
- The first 1800 points are not paid.

- Example: Suppose you earn 1980 points – which is 180 above 1800.
  Thus, you would receive 180*0.05€ = 8.00€ for the main task.

back START

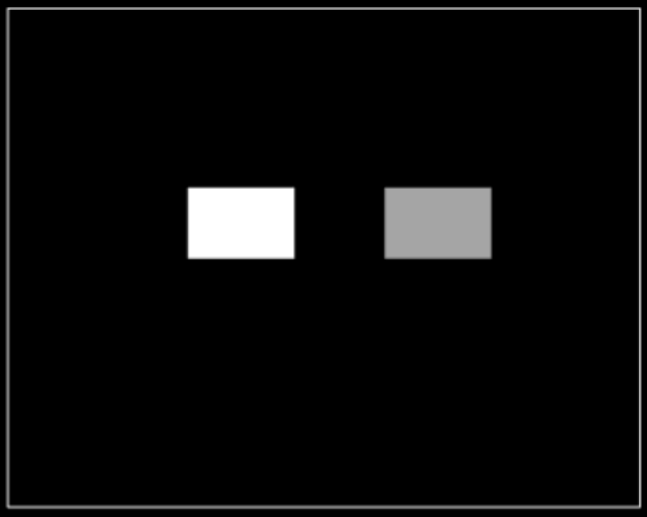
Instructions and Quiz: Patterns Task



Your choice options.

What you see on the right is similar to what you will see during the task.

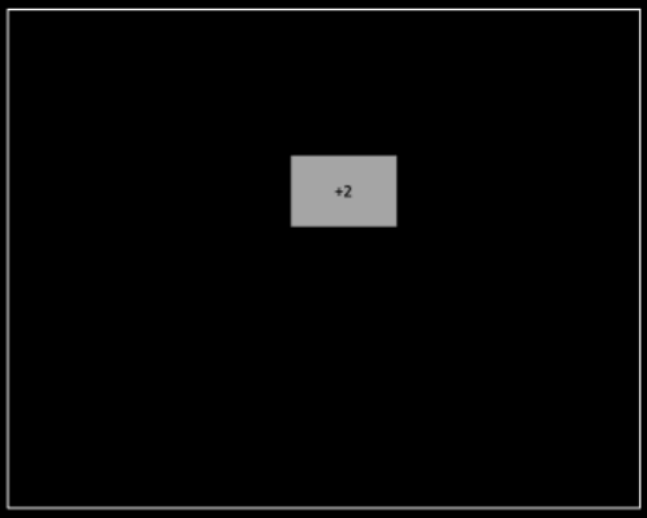
In each round, you will see two images. You will have to choose one of them. In the task, these rectangles will be different images. In these instructions they are instead shades of gray.



Feedback screen

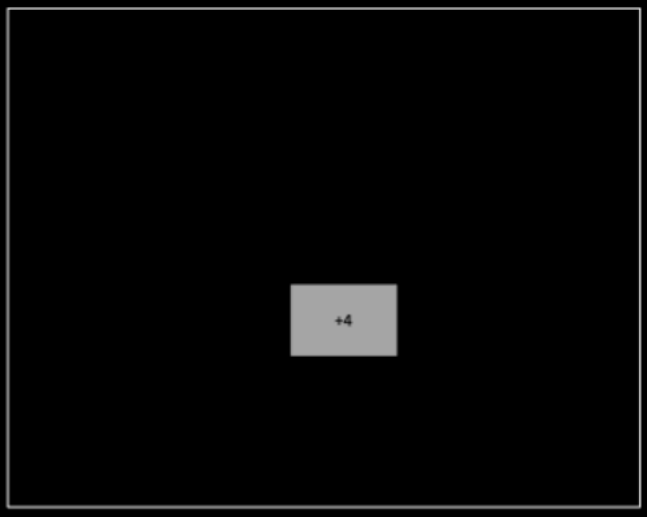
Each image will give you 2 rewards.

The first reward will be displayed immediately after your choice (top rectangle).



Feedback screen

The second reward will be displayed in the following round after your next choice (bottom rectangle).



Feedback screen

The rewards depend on the image you choose and on chance. Each image has a different average reward, but there is some randomness so that each reward can vary by up to 3 points.

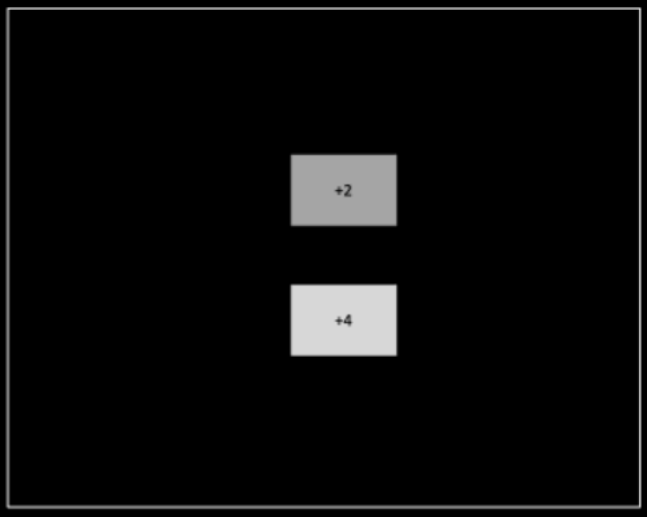
The more points you earn in the task, the higher your payment will be.

Feedback screen

This is what a feedback screen would really look like. The top rectangle (dark gray) is the one you just chose. On it you can see the first reward that this image has given you.

The bottom rectangle (light gray) is the one you chose in the previous round. On it you can see the second reward that this image has given you.

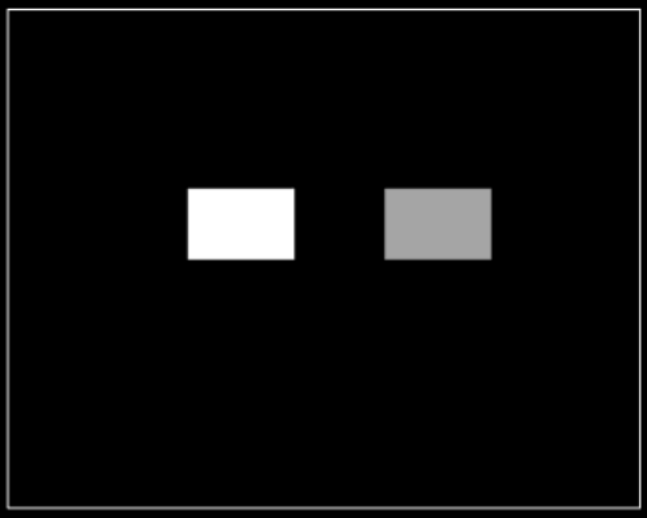
It's worth noting that the first and second reward from an image are usually different.



Making your choice

You make your choice by pressing the left arrow key (for the left image) or the right arrow key (for the right image).

If you do not decide in 3 seconds, the computer will randomly choose one image for you. Each time you fail to decide within 3 seconds, we will **subtract 5 points** from the total number of points you earned in the experiment.



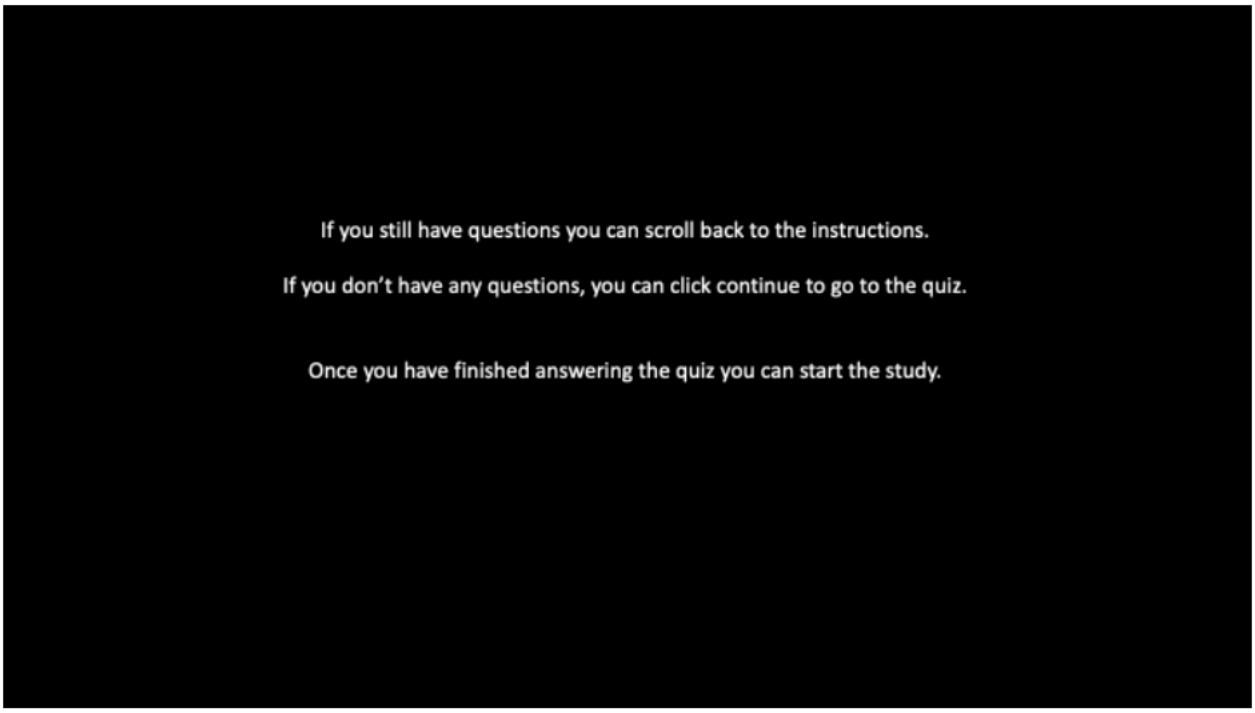
Your payment for the study will be calculated as follows:

- You will receive **4 cents** for each point you earn after you have achieved 1505 points.
- The first 1505 points are not rewarded.

- Example: you gather 1805 points – that is 300 points above 1505.
- You would therefore receive 300 x 0.04 = $12.00.

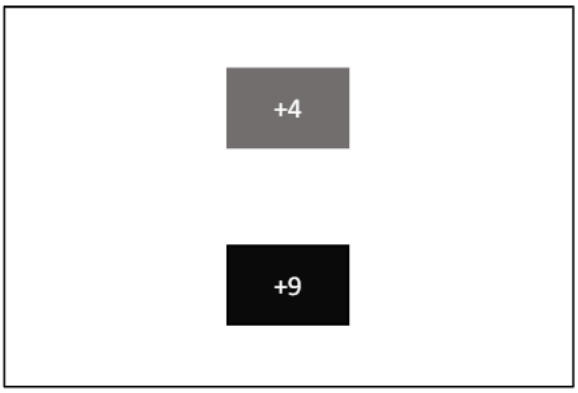
Summary

- You will choose between **2 images** in each round.
- Each image will bring **2 rewards**, in points. Points are converted to cash at the end.
- The first and second reward for an image are usually different.
- The average reward for each image is different.
- You have **3 seconds** to make each decision, by pressing the **left or right arrow keys**.

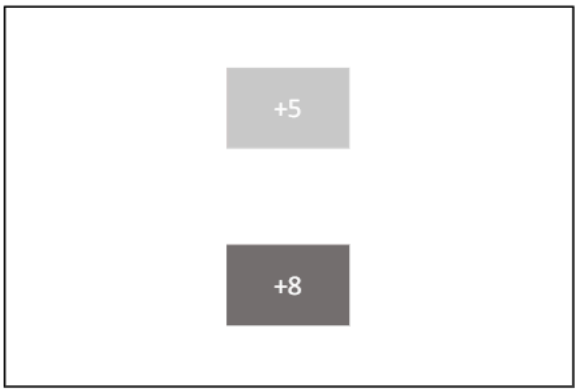

If you still have questions you can scroll back to the instructions.
If you don't have any questions, you can click continue to go to the quiz.
Once you have finished answering the quiz you can start the study.

Consider the following example displaying choices and payoffs from two consecutive rounds of the task.

Round 31 – Feedback screen



Round 32 – Feedback screen



Question 1: What does the top number (+4) in Round 31 represent?

- **The first payoff for the dark gray square.**
- The payoff is not relevant to the task.
- The first payoff for the black square.
- The second payoff for the dark gray square.

Question 2: What does the bottom number (+9) in Round 31 represent?

- The payoff is not relevant to the task.
- The first payoff for the black square.
- **The second payoff for the black square.**
- The first payoff for the dark gray square.

Question 3: What is your total payoff for Round 31?

- 4
- 9
- 11
- **13**

Question 4: What is the total payoff of the dark gray square from Round 31 and Round 32?

- **12**
- 16
- 13

Question 5: What was chosen in Round 32?

- The dark gray square.
- **The light gray square.**

Question 1

In round 31, it can be inferred from the feedback screen that the dark gray square was chosen in that round and the white square was chosen in the previous round.

Recall that each image will give you 2 payoffs; one immediately after your choice (always the top image) and one in the next round (always the bottom image). Therefore, the number **+**4 on the top image (the dark gray square) represents the **first payoff for the dark gray square**.

Question 2

In round 31, it can be inferred from the feedback screen that the dark gray square was chosen in that round and the white square was chosen in the previous round.

Recall that each image will give you 2 payoffs; one immediately after your choice (always the top image) and one in the next round (always the bottom image). Therefore, the number **+**9 on the bottom image (the white square) represents the **second payoff for the white square**.

Question 3

Your total payoff in points for a round is the sum of the payoffs for that round. In this case, for round 31 you got:

- **+** 4 as the first payoff for the dark gray square chosen in that round.
- **+** 9 as the second payoff for the white square chosen in the previous round.

Your total payoff is therefore: 4 **+** 9 = **13.**

Question 4

Recall that each image will give you 2 payoffs; one immediately after your choice (always the top image) and one in the next round (always the bottom image). The dark gray square gives you **+** 4 in round 31 and **+** 8 in round 32. Therefore, the total payoff for the dark gray square is **12.**

Question 5

Recall that each image will give you 2 payoffs; one immediately after your choice (always the top image) and one in the next round (always the bottom image).

In round 32, it can be inferred from the feedback screen that the **light gray square** was chosen in that round and the dark gray square was chosen in the previous round.