Risk paper

Samuel Zorowitz[1], Yael Niv[1,2]

[1]Princeton Neuroscience Institute, Princeton University, USA
[2]Department of Psychology, Princeton University, USA

Reinforcement learning describes a cognitive and statistical framework for understanding how humans and other animals learn to make reward-maximizing (or punishment-minimizing) decisions (Niv, 2009; Sutton and Barto, 2018). At the core of reinforcement learning is the prediction error, which is the difference between an observed and expected outcome. A positive prediction error occurs when an outcome is better than anticipated, whereas a negative prediction error occurs when an outcome is worse than anticipated. For example, a person might experience a positive prediction error if, after choosing to dine at an unfamiliar restaurant, the food is tastier than expected; so too, a person might experience a negative prediction error after making the same choice if the food was worse than expected. Through the integration of prediction errors, humans can learn which choices lead to desirable outcomes and are thus worth repeating in the future.

An important source of individual variation in reinforcement learning is differential sensitivity to signed prediction errors; that is, learning more from either positive or negative prediction errors. Differential sensitivity to signed prediction errors has been implicated in individual differences in a number of cognitive processes, including approach-avoidance decision-making (Frank et al., 2007), risk sensitivity (Niv et al., 2012), optimism and confirmation biases (Lefebvre et al., 2017; Palminteri and Lebreton, 2022), and the formation and maintenance of self-relevant beliefs (Garrett et al., 2018). Differential sensitivity has also been hypothesized or implicated to play a role in natural variation in psychiatric syndromes such as depression (Bennett and Niv, 2020), anxiety (Aylward et al., 2019; Cavanagh et al., 2019), bipolar mania (Ossola et al., 2020), and attention-deficit/hyperactivity disorder (Cockburn and Holroyd, 2010). Recent research has provided neural circuit-level evidence for differential sensitivity in the form of dopaminergic subpopulations with differential sensitivity to the sign (and magnitude) of prediction errors (Dabney et al., 2020).

Differential sensitivity to signed prediction errors is conventionally measured by fitting a reinforcement learning model to an individual's choice behavior (e.g., on a cognitive task), where the model includes separate learning rates for positive and negative prediction errors, and the

asymmetry in the magnitudes of these parameters indicates the degree of differential sensitivity. There exists, however, a critical challenge to successfully measuring differential sensitivity in such a way. Specifically, Katahira (2018) demonstrated that unmodeled autocorrelation in choice data (i.e., the dependence of current choice on past choice, independent of the value of the choice options) may severely bias estimates of asymmetric learning rates. That is, choice perseveration (i.e., the tendency, all else being equal, to repeat a previous choice), if unaccounted for in an asymmetric learning rates model, can yield positive asymmetries, giving the impression of differential sensitivity to signed prediction errors where there is objectively none. In recent follow-up work, Sugawara and Katahira (2021) demonstrated that this issue occurred in practice in empirical reinforcement learning experiments. Specifically, they showed that purportedly positive asymmetries could be explained away by the inclusion of additional model parameters that accounted for choice autocorrelation. In sum, choice autocorrelation poses a serious threat to the validity of studies investigating differential sensitivity to signed prediction errors.

This issue raises at least two key questions. The first is, what is the source of choice autocorrelation? One possibility is that participants simply perseverate on past choices; that is, for heuristic reasons, participants simply tend to repeat past choices. Here we will raise an alternative explanation. When we fit reinforcement learning models to participants' data, we typically assume that learning rates are constant; that is, we assume participants remain equally sensitive to prediction errors throughout the course of a task. In experiments with stationary reward distributions (i.e., when reward contingencies are unchanging), this assumption is almost certainly wrong. In stationary reward environments, it is normative to decrease learning rates with successive feedback. This is because, if an agent knows an environment is unchanging, later feedback provides no new information above and beyond early feedback. There is considerable evidence that humans (and other animals) adjust their learning rates adjust with successive feedback (Camerer and Hua Ho, 1999; Craig et al., 2011; Nassar et al., 2012; Pearce and Hall, 1980). Importantly, when learning rates are reduced, agents will continue to choose what they have already been choosing; that is, they will exhibit choice perseveration. Though the link between nonstationary learning rates and choice perseveration has been suggested elsewhere (Harada, 2020; Palminteri, 2022), the possibility of it biasing asymmetric learning rates has not of yet been thoroughly explored.

The second question is simply what to do about it. Katahira (2018) proposed addressing this issue through additional modeling. Specifically, they suggest researchers interested in asymmetric learning rates fit a hybrid reinforcement learning, involving the addition of a choice kernel to the model. However, this solution is not without its own methodological and interpretational issues (a point we return to later). Taking a step back, this solution is suboptinal insofar that it treats the problem after its occurred. Here, we propose to instead treat the issue earlier, at the experimental design stage. As we will show, by better understanding the source of choice autocorrelation, we can better design tasks to be robust against it.

1 Studies 1a & 1b

In the first set of studies, we investigated two experimental paradigms that have used to study differential sensitivity to signed prediction errors during reinforcement learning: the two-alternative forced choice (2AFC) task (Chambon et al., 2020; Lefebvre et al., 2017; Sugawara and Katahira, 2021) and the risk sensitivity task (RST; Arkadir et al., 2016; Niv et al., 2012; Rosenbaum et al., 2022). The goals for both studies were threefold. First, we reanalyzed archival datasets to determine if empirical choice behavior on these tasks were consistent with a nonstationary (e.g., decreasing) learning rates model. Second, we explored whether nonstationary learning could, if unmodeled, masquerade as asymmetric learning rates on these tasks. To do so, we simulated choice behavior on both tasks under a nonstationary learning model using parameters informed by the archival dataset analyses. Then, we fit an asymmetric learning rate model to the simulated data and quantified the degree of bias in the estimated parameters. Finally, we performed a model recoverability analysis to find out if choice behavior under three qualitatively-distinct reinforcement learning models (i.e., asymmetric learning rates, nonstationary learning rates, perseverative choice) are dissociable on these tasks. This final analysis is crucial for individual difference studies employing these tasks because, if these models are confusable, then a researcher cannot confidently conclude what choice strategy most accurately describes any given participant.

1.1 Methods

1.1.1 Experimental paradigms

2-alternative forced choice (2AFC) task. On every trial, participants are presented with two cues, each associated with a stationary (i.e., unchanging) reward probability. For example, one cue might return reward (`+1` point) with a 75% chance and otherwise return a non-reward (`0` points) or punishment (`-1` points). The cues may have the same (e.g., 50%/50%) or different (e.g., 75%/25%) reward probabilities. Participants are not informed about the cue-outcome contingencies, but must learn them from experience. Typically participants are instructed, in order to maximize their earnings, they should learn and choose those cues that most consistently return rewards.

Risk sensitivity task. The RST is a variant of the 2AFC task. In the task, there are four cue types. Three of these are each associated with a deterministic outcome: a certain non-reward (e.g., `0` points), a certain moderate reward (e.g., `+1` points), or a certain large reward (e.g., `+2` points). The fourth cue type (i.e., risky) is associated with a stationary reward probability, returning a large reward with 50% probability and a non-reward otherwise. Participants are not informed about the cue-outcome contingencies, but must learn them from experience. On the majority of trials, participants must choose one of the pseudo-randomly

paired cues. (The minority of trials are forced choice trials, in which participants are made to choose one particular cue.) Of special interest are trials in which participants are required to choose between the certain moderate-reward cue and risky cue. In expectation, these cues are of equal value. However, if a participant is differentially sensitive to positive (negative) prediction errors, then through learning, they should come to favor the risky (certain) cue.

Archival datasets

Lefebvre et al. (2017). A total of N=85 in-lab participants completed a 2AFC task. The task involved a four pairs of cues with differing reward probabilities (i.e., 75%/25%, 25%/75%, 25%/25%, 75%/75%). For all cues, the better of the two possible outcomes was winning money. For some participants, the worse of the two outcomes was non-reward, whereas for others it was losing money. Participants completed four blocks of 24 trials, so that 96 trials were available per participant.

Fontanesi et al. (2019). A total of N=89 in-lab participants completed multiple blocks of a 2AFC task that manipulated both feedback valence (reward, punishment) and feedback type (factual, counterfactual). Only the blocks involving factual feedback were used in the current analyses. For all participants, the cue outcome probabilities were 75%/25%. In reward blocks, the better outcome was winning either points or money and the worse outcome was a non-reward. In punishment blocks, the better outcome was a non-outcome (i.e., avoiding a loss) and the worse outcome was losing either points or money. Participants completed either four blocks of 20 trials or six blocks of 24 trials, so that either 80 or 144 trials were available per participant.

Chambon et al. (2020). A total of N=102 in-lab participants completed multiple blocks of a 2AFC task that manipulated feedback type (factual, counterfactual), choice agency (free choice, fixed choice), and action type (cue choice, go/no-go response). Only the blocks involving free choices between pairs of cues with factual feedback were used in the current analyses, restricting the sample to N=30 participants. Half of the included blocks involved 70%/30% reward probabilities, whereas the other half involved 50%/50% reward probabilities. In all blocks, the better outcome was winning points and the worse outcome was losing points. Participants completed six blocks of 20 trials, so that 120 trials were available per participant.

Sugawara and Katahira (2021). A total of N=143 online participants completed multiple blocks of a 2AFC task that manipulated feedback type (factual, counterfactual) and reward type (stationary, volatile). Only the blocks involving factual feedback and stationary rewards were used in the current analyses. Half of the included blocks involved 75%/25% reward probabilities, whereas the other half involved 50%/50% reward probabilities. For all blocks, the

better outcome was winning points, whereas the worse outcome was losing points. Participants completed eight blocks of 24 trials, so that there were 192 trials available per participant.

Radulescu et al. (2020). A total of N=563 online participants completed 176 trials of the RST. On 126 trials, participants had to choose between pairs of cues (the remaining trials were fixed choice trials). Of these, 30 involved a choice between the certain moderate-reward cue and risky cue. The order of trials were randomized across participants.

Bennett et al. (unpublished). A total of N=150 online participants completed 156 trials of the RST. On 120 trials, participants were required to choose between pairs of cues (the remaining trials were fixed choice trials). Of these, 36 involved a choice between the certain moderate-reward cue and risky cue. The order of trials were randomized across participants.

1.1.2 Exclusion criteria

For the 2AFC task datasets, we excluded participants who chose the reward-maximizing cue (in blocks involving unequal reward probabilities) on fewer than 60% of trials. This resulted in the removal of 120 out of 347 total participants (34.6%). We note that the majority of excluded participants (N=70, 58.3%) were from Sugawara and Katahira (2021). Of the four archival 2AFC datasets, only Sugawara and Katahira (2021) recruited their participants online, which may explain the ostensibly worse data quality. An additional six participants (1.7%) were excluded for missing responses on more than 5% of trials. This left the data from 221 participants (63.7%) available for analysis.

For the RST datasets, we excluded participants who failed to choose the better cue on at least 80% of trials involving a dominant choice option (e.g., certain `+2` points or certain `0` points). This resulted in the removal of 291 out of total 713 participants (40.8%). We note that almost all of these excluded participants (N=270, 92.3%) were from Radulescu et al. (2020). Radulescu et al. (2020) collected these data before the advent of more recent data quality protocols (e.g., Litman et al., 2020), which may explain the ostensibly worse data quality. This left the data from 422 participants (59.2%) available for analysis.

1.1.3 Reinforcement learning models

We compared three reinforcement learning models of choice behavior on the 2AFC task and RST. The basis of all three models is the canonical Q-learning model (Sutton and Barto, 2018), under which the dynamics of choice and learning are governed two rules. Under the model, the probability that an agent chooses the first of a pair of cues, a_1, is defined as:

$$p(y = a_1) = \text{logit}^{-1}\left(\beta \cdot [Q(s, a_1) - Q(s, a_2)]\right) \tag{1}$$

where $Q(s, a_1)$ and $Q(s, a_2)$ are the current value estimates of the two choice options, and β is the choice sensitivity (inverse temperature) parameter. Choice sensitivity controls an agent's sensitivity to the difference in value between the choice options. As $\beta \to \infty$, a participant is increasingly deterministic in selecting the highest value option; as $\beta \to 0$, a participant is increasingly stochastic in their choice.

In turn, the values of a choice option, $Q(s, a)$, is learned according to the delta rule:

$$Q(s, a) \leftarrow \eta \cdot \delta \tag{2}$$

where δ is the reward prediction error, and η is a learning rate defined in the range $\eta \in [0, 1]$. The reward prediction error is defined in turn as:

$$\delta = r - Q(s, a) \tag{3}$$

where r is the observed outcome on a particular trial. As previously discussed, the reward prediction error is the difference between the observed and expected outcome. The learning rate, η, governs the degree to which value estimates reflect more recent outcomes. As $\eta \to 1$, value estimates increasingly reflect the most recent outcomes; as $\eta \to 0$, value estimates reflect the integration over longer histories of outcomes.

Asymmetric learning rates model. The first model we investigated is the asymmetric learning rates model, in which there are separate learning rates for positive and negative prediction errors. Specifically, the learning rule for the model is defined as:

$$Q_k \leftarrow \begin{cases} \eta_+ \cdot \delta & \text{if } \delta > 0 \\ \eta_- \cdot \delta & \text{if } \delta < 0 \end{cases} \tag{4}$$

where η_+ and η_- are independent learning rate parameters that control Q-value updates for positive and negative prediction errors, respectively. The choice rule for this model is as defined as in (1).

Crucially, the two learning rates are permitted to vary in magnitude. When $\eta_+ > \eta_-$, an agent is more sensitive to positive prediction errors than negative predictions (and vice-versa). The degree of differential sensitivity to signed prediction errors can be quantified via the asymmetry index, κ:

$$\kappa = \frac{(\eta_+ - \eta_-)}{(\eta_+ + \eta_-)} \tag{5}$$

where $\kappa > 0$ indicates greater sensitivity to positive prediction errors and $\kappa < 0$ indicates greater sensitivity to negative prediction errors. We use the asymmetry index because it standardizes the difference in learning rates across disparate average learning rate magnitudes.

Nonstationary learning rates model. The second model we investigated is the nonstationary learning rates model, in which the learning rate evolves with successive prediction

errors. Specifically, the learning rule for this model is defined as:

$$\eta_t = \text{logit}^{-1}\left(a_0 + a_1 \cdot t\right) \tag{6}$$

where a_0 is an intercept parameter, controlling the initial learning rate; a_1 is a slope parameter, controlling how the learning rate changes with each new prediction error; and t is the number of times an agent has previously chosen a particular cue (initialized at $t = 0$). We offer three remarks about this learning rule. First, the use of the logistic function in (6) ensures that the learning rate is in the range $\eta_t \in [0, 1]$. Second, when $a_1 = 0$, the learning rate is constant and the model reduces to the canonical Q-learning model. Third, a_1 is permitted to be positive or negative; that is, the learning rate is permitted to grow ($a_1 > 0$) or decay ($a_1 < 0$) with every new prediction error. The choice rule for this model is also as defined as in (1).

Perseverative choice model. The third model we investigated is the impulse perseverative choice model, in which choice on the current trial is partially determined by choice on the previous trial. Specifically, the choice rule for this model is defined as:

$$p(y = a_1) = \text{logit}^{-1}\left(\beta_1 \cdot [Q(s, a_1) - Q(s, a_2)] + \beta_2 \cdot [C_1(s, a_1) - C(s, a_2)]\right) \tag{7}$$

where $C_1(s, a_1)$ and $C_2(s, a_2)$ are binary indicators denoting whether a particular cue was chosen on the previous trial, and β_2 is an autoregressive choice parameter. If $\beta_2 > 0$, an agent is more likely to repeat their choice from the previous trial; if $\beta_2 < 0$, they are more likely to switch to the other choice. The learning rule for this model is the same as in (2) and (3).

1.1.4 Analyses

In line with our goals, we performed three analyses for each task separately. In the first analysis, we fit the nonstationary learning rates model to each participant's choice data from the six archival datasets. The primary motivation for this analysis was to determine if participants' choice behavior was, on average, consistent with nonstationarity in learning rates (i.e., $a_1 \neq 0$). Three of the 2AFC task datasets involved blocks with with equal reward probabilities (e.g., 50%/50%). We did not want to assume that learning dynamics would be equivalent in blocks with and without a clear reward signal. As such, we fit separate learning rule parameters (i.e., a_0, a_1) for these blocks.

In the second analysis, we explored whether nonstationary learning, if unaccounted for, can masquerade as asymmetric learning. To do so, we simulated choice behavior on both tasks under the nonstationary learning model under three conditions — stationary learning, moderate nonstationarity, and large nonstationarity — where the data-generating parameters were informed by the archival data analysis. For the 2AFC task, artificial agents completed six blocks of 24 trials (144 total trials), where half the blocks involved either unequal (75%/25%) or equal (50%/50%) reward probabilities. For the RST, artificial agents completed 36 free

choice trials (i.e., choosing between the certain, moderate-reward cue and risky cue) and 36 forced choice trials (i.e., choosing the risky cue). These task configurations were chosen to mimic realistic experimental conditions; that is, to resemble the experimental designs from the archival datasets. For each task and nonstationarity condition, 500 unique datasets were simulated. The asymmetric learning rates and perseverative choice models were then fit to each simulated dataset. We then inspected if unmodeled nonstationarity would manifest as asymmetric learning rates, as measured by the asymmetry index (κ), and as choice perseveration, as measured by the choice autocorrelation parameter (β_2).

In the final analysis, we investigated the recoverability of the three reinforcement learning models on both tasks. Specifically, we simulated choice behavior on the 2AFC task and RST, using the task configurations described above, under the asymmetric learning rates, nonstationary learning rate, and perseverative choice models. We simulated data for each model across 1089 unique combinations of parameters (Table S1), each sampled twice, yielding a total of 2178 datasets per model. We then fit the same three models to every simulated dataset and calculated the proportion of times the best-fitting model was also the true data-generating model. Because each model is equally complex (i.e., three free parameters per model fit), we compared the fit of each model directly using the marginal log-likelihood.

In all analyses, the reinforcement learning models were fit to data (empirical or simulated) using *maximum a posteriori* (MAP) estimation using the L-BFGS algorithm as implemented in cmdstan (v2.30; Carpenter et al., 2017). We elected not to use more sophisticated model-fitting procedures (e.g., Hamiltonian Monte Carlo) due to would-be computational burden of fitting those models to the large number of simulated datasets. Our choices of prior for each model are detailed in the Supplementary Materials. Briefly, we specified diffuse and uninformative priors in order to avoid biasing parameter estimation.

1.2 Results: Study 1a

1.2.1 Empirical choice behavior on the 2AFC task is consistent with nonstationary (i.e., decreasing) learning rates.

Posterior predictive checks showed the nonstationary learning rates model provided an acceptable fit to the archival 2AFC task datasets (Figure S1). Moreover, permutation paired-sample t-tests indicated that neither the initial learning rate (a_0: $t = 1.560$, $p = 0.118$) nor the learning rate slope (a_1: t = -1.414, p = 0.160) significantly differed across equal and unequal reward probability blocks. Therefore, we averaged across these parameters by block for all subsequent analyses.

Group-averaged estimates of trial-by-trial learning rates are presented in Figure 1a. In every dataset, the models indicated decreased learning from successive prediction errors on average. Collapsing across datasets, the grand-average learning rate slope was negative and significantly

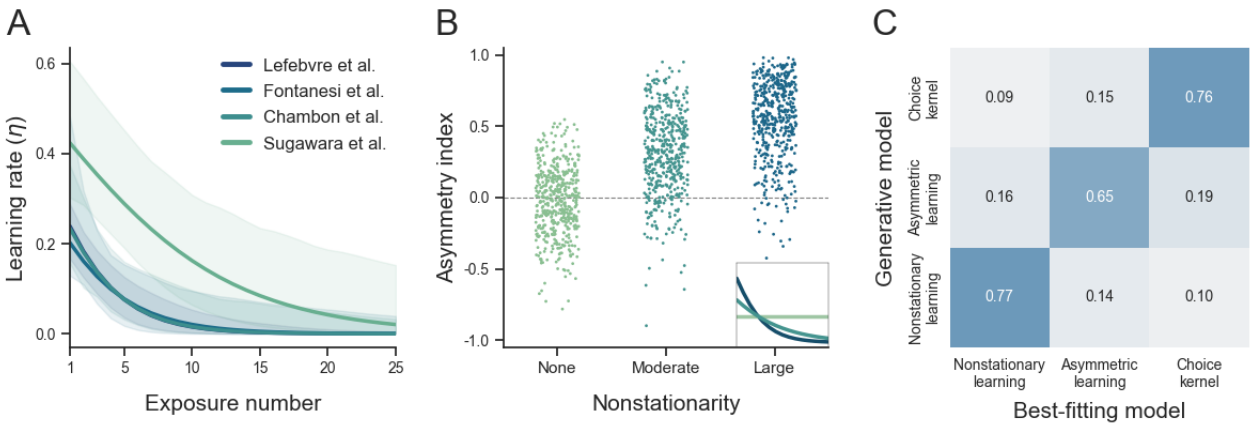

Figure 1: (A) Model-predicted trial-by-trial learning rates by archival dataset. On average, the participants in each dataset showed evidence of nonstationary (i.e., decreasing) learning rates. Shaded regions correspond to 95% bootstrapped confidence intervals. (B) Model-predicted asymmetry indices (κ) as a function of unmodeled nonstationary learning. Inset: The latent learning rate curves underlying each condition. (C) Confusion matrix for the model recoverability analysis for the 2AFC task. Values indicate the proportion of times a data-generating model (y-axis) was also the best-fitting model (x-axis).

different than zero ($\bar{a}_1$ = -0.606, t = -8.784, $p < 0.001$). In conjunction with the grand-average initial learning rate ($\bar{a}_0 = -0.429$), the averaged model parameters predict that learning rates reach a lower asymptote ($\eta < 0.01$) after approximately eight prediction errors following choice of a particular cue. In short, the results are consistent with the hypothesis that participants exhibit nonstationary learning on the 2AFC task.

1.2.2 Unmodeled nonstationary learning can masquerade as asymmetric learning rates on the 2AFC task

Next, we investigated whether nonstationary learning rates, if unaccounted for, can bias asymmetric learning rates. We simulated choice data on the 2AFC task under three parameterizations of the nonstationary learning rates model: stationary learning ($a_0 = -1.821$, $a_1 = 0.000$), corresponding approximately to the 25th percentile of observed parameters from the archival data; moderate nonstationary learning ($a_0 = -1.245$, $a_1 = -0.179$), corresponding to the 50th percentile of parameters; and large nonstationary learning ($a_0 = -0.677$, $a_1 = -0.353$), corresponding to the 75th percentile of observed parameters. Across all simulations, we assumed one single inverse temperature ($\beta_1 = 7.5$). The asymmetric learning rates model was then fitted to each choice dataset.

The results of the simulations are presented in Figure 1b. Under a stationary (i.e., constant) learning rate, the asymmetry index is centered approximately at zero ($\bar{\kappa}$ = -0.022, sd = 0.234). Under moderate nonstationarity, however, a positive bias in the asymmetry index is observed ($\bar{\kappa}$ = 0.307, sd = 0.292). Under large nonstationarity, the bias further increases ($\bar{\kappa}$ = 0.530, sd = 0.312). The results unequivocally demonstrate that nonstationary learning, if unaccounted

for, can masquerade as positively-biased asymmetric learning rates (where the degree of the bias is dependent on the severity of learning rate decay). A similar pattern of results was observed for the perseverative choice model; there, larger degrees of nonstationarity in learning rates led to increasingly positive choice autocorrelation parameters (Figure S4a).

It is worth pausing to consider why this occurs. Consider a single block of the 2AFC task with 75%/25% cue outcome probabilities. At the start of the block, a reward-sensitive nonstationary learning agent will, through trial-and-error, learn to choose the better cue. As the agent continues to choose the cue, its learning rate will decrease, and it will consequently grow increasingly insensitive to every prediction error (both positive and negative). As such, it will become insensitive to the occasional negative prediction error (i.e., following a non-reward, expected on 25% of trials) that might otherwise cause it to choose the other cue; that is, a nonstationary learning agent on the 2AFC task is expected to perseverate on a choice option despite occasional negative feedback. As a consequence, if an asymmetric learning rates model is fit to the data, it will account for this ostensible insensitivity to negative prediction errors by increasing the positive learning rate relative to the negative learning rate. (More straightforwardly, if the perseverative choice model is fit to the data, it will account for the perseveration in choice by increasing the choice autocorrelation parameter.)

Before continuing, we highlight two moderators of this result. First, the degree of bias in asymmetric learning rates is in part a function of the number of trials in a block of the 2AFC task. As the number of trials increases, so too does the positive bias (Figure S3a). The reason why is straightforward: with longer blocks, more trials are sampled at asymptotic learning (i.e., where a nonstationary learning agent is least sensitive to feedback). Second, the degree of bias in asymmetric learning rates is also affected by lower asymptote of the learning curve. Importantly, the bias is attenuated but not abolished if the learning rates asymptote at a non-zero value (Figure S3b).

1.2.3 Distinct reinforcement learning models are confusable on the 2AFC task

The results of the model recoverability analysis are summarized in Figure 1c. While far from chance-levels, true-positive rates did not exceed 80%. Of greatest concern is the 65% true-positive rate for the asymmetric learning rates model. That is, in more than one third of cases, a researcher might mischaracterize a participant with true differential sensitivity to signed prediction errors as instead exhibiting nonstationary learning or choice perseveration.

In summary, typical versions of the 2AFC task are not well suited for studying individual differences in asymmetric learning rates. We turn our attention next to the RST to see if it fares any better.

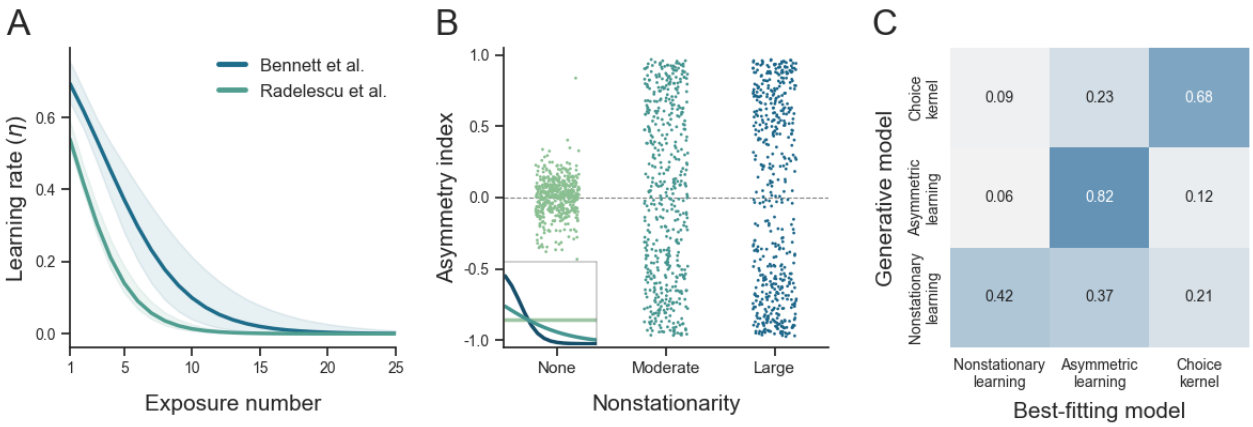

Figure 2: (A) Model-predicted trial-by-trial learning rates by archival dataset. On average, the participants in each dataset showed evidence of nonstationary (i.e., decreasing) learning rates. Shaded regions correspond to 95% bootstrapped confidence intervals. (B) Model-predicted asymmetry indices (κ) as a function of unmodeled nonstationary learning. Inset: The latent learning rate curves underlying each condition. (C) Confusion matrix for the model recoverability analysis for the RST. Values indicate the proportion of times a data-generating model (y-axis) was also the best-fitting model (x-axis).

1.3 Results: Study 1b

1.3.1 Empirical choice behavior on the RST is consistent with nonstationary (i.e., decreasing) learning rates.

Posterior predictive checks indicated that the nonstationary learning rates model provided an acceptable fit to the archival RST data (Figure S2). Group-averaged estimates of trial-by-trial learning rates are presented in Figure 2b. In both datasets, the models indicated decreased learning from successive prediction errors on average. Collapsing across datasets, the grand-average learning rate slope was negative and significantly different than zero ($\bar{a}_1$ = -0.628, t = -13.300, $p < 0.001$). In conjunction with the grand-average initial learning rate ($\bar{a}_0 = 0.644$), the averaged model parameters predict that learning rates reach a lower asymptote ($\eta < 0.01$) after approximately ten prediction errors following choice of a particular cue. In short, the results are consistent with the hypothesis that participants exhibit nonstationary learning on the RST.

1.3.2 Unmodeled nonstationarity masquerades as asymmetric learning rates

Next, we investigated whether nonstationary learning rates, if unaccounted for, can bias asymmetric learning rates. We simulated choice data on the RST under three parameterizations of the nonstationary learning rates model: stationary learning ($a_0 = -0.837$, $a_1 = 0.000$), corresponding approximately to the 20th percentile of observed parameters from the archival data; moderate nonstationary learning ($a_0 = 0.266$, $a_1 = -0.361$), corresponding to the 50th

percentile of parameters; and large nonstationary learning ($a_0 = 1.674$, $a_1 = -0.717$), corresponding to the 80th percentile of observed parameters. Across all simulations, we assumed one single inverse temperature ($\beta_1 = 7.5$). The asymmetric learning rates model was then fitted to each choice dataset.

The results of the simulations are presented in Figure 2b. Under a stationary (i.e., constant) learning rate, the asymmetry index is centered approximately at zero ($\bar{\kappa} = 0.007$, sd = 0.132). Under moderate nonstationarity, however, both positive and negative biases in the asymmetry index are observed ($|\bar{\kappa}| = 0.536$, sd = 0.286). Under large nonstationarity, the bias appears almost bimodal ($|\bar{\kappa}| = 0.620$, sd = 0.261). The results demonstrate that nonstationary learning, if unaccounted for, can masquerade as either positively- or negatively-biased asymmetric learning rates (where the magnitude of the bias is dependent on the severity of learning rate decay). However, a divergent pattern of results was observed for the perseverative choice model; there, larger degrees of nonstationarity in learning rates led only to increasingly positive choice autocorrelation parameters (Figure S4b).

The explanation for this result follows the same line of reasoning as for the 2AFC task. If, by chance, a nonstationary learning agent obtains a greater than average number of rewards from the risky cue, it will be likely to choose that it again in the future. However, with each choice, it grows increasingly insensitive to feedback (both positive and negative prediction errors), therefore making it less likely to switch to an alternative cue following the occasional non-reward. If this agent's choice behavior is then fit with the asymmetric learning rates, the models will compensate for this pattern of behavior by increasing the positive learning rate relative to the negative one. In this way, the explanation is identical as that for the 2AFC task. In contrast, however, the opposite pattern is also possible on the RST. A nonstationary learning agent that experiences an early series of non-rewards from the risky cue is, by virtue of its decreasing sensitivity to future feedback, likely to continue to avoid it in spite of the occasional positive prediction error. Thus, the asymmetric learning rates model will account for this behavior by decreasing the positive learning rate relative to the negative one. Interestingly, because both patterns of behavior involve choice perservation, the autocorrelation parameter of the impulse choice kernel model will always be positively-biased (Figure S4b).

1.3.3 Distinct learning models are confusable on the RST

The results of the model recovery analysis are summarized in Figure 2c. In contrast, the model recovery rate for the asymmetric learning rates model (82%) was notably improved compared to the 2AFC task, though the recovery rate choice kernel model was worse. However, the true-positive rate for the nonstationary learning rates model was very poor and almost at chance levels (37%). In other words, choice behavior produced by a nonstationary learning rates model on the RST is easily explained by either the asymmetric learning rates or choice kernel model. In sum, model recovery on the RST is less than optimal for studying investigating asymmetric learning rates.

1.4 Interim discussion

The results of studies 1a and 1b...

First, we do not believe that the current results should be viewed as reason to cast doubt on previous empirical investigations of asymmetric learning rates. That an experiment may be confounded does not necessitate it is, in fact, confounded. There is ample reason to believe that the asymmetric learning rates previously observed in the datasets we reanalyzed here could not be entirely explained away by unmodeled nonstationary learning. For example, Palminteri (2022) recently re-analyzed multiple 2AFC task datasets (including some of the datasets analyzed here) using a hybrid asymmetric learning rates and perseverative choice model (i.e., a model that should at least partially account for the biasing effects of choice autocorrelation that stems from nonstationary learning). They found that, even after controlling for perseverative choice, asymmetric learning persisted (albeit attenuated) at the group-level. Separately, Niv et al. (2012) observed strong coupling between participants' preferences for the risky cue on the RST and fMRI BOLD activation in the nucleus accumbens, thereby providing an independent neural source of evidence for the existence differential sensitivity to signed prediction errors. To summarize, though our results call into question the utility of the 2AFC and RST paradigms for studying individual differences in asymmetric learning rates, they do not necessarily invalidate the findings of all previous studies who did so.

Second, we should note two points where our methods or results diverge from previous studies. Katahira (2018) makes clear that the one-trial-back perseverative choice model (i.e., the model used in the preceding analyses) is unable to fully account for more graded forms of choice autocorrelation, and therefore suggests researchers instead incorporate the full, two-parameter choice kernel into their reinforcement learning models. We elected to use the former model, however, because in our experience the two-parameter model is liable to inaccurately explain away reward-guided choice (e.g., when the latent data-generating model is not perseverative; Figure S5). As such, the addition of the full choice kernel to a reinforcement learning can lead to serious parameter interpretation issues. Separately, Sugawara and Katahira (2021) investigated the recoverability of the asymmetric learning rates and perseverative choice models on the 2AFC task and found it highly capable of discriminating between choice behavior generated under those two models. However, their analyses involved 960 total trials per simulated agent; that is, they studied model recoverability using more than five times more trials than what is typically collected empirically. We therefore conclude that, although the 2AFC task can exhibit good model recoverability, it is unlikely to under more realistic conditions.

2 Study 2

To address the issues we identified in the 2AFC and RST paradigms, we designed a novel 2AFC task — the Double-or-Nothing (DoN) task — that takes inspiration from the RST. A schematic

of a single trial of the DoN task is presented in Figure 3a. On every trial, participants are presented with two cues: a certain cue and a risky cue. The certain cue is associated with a deterministic moderate reward (`+5` points), which is made explicit to participants. The risky cue is associated with a static probability of returning a large reward (`+10` points) and otherwise returns a non-reward (`0` points). There are four types of risky cues in the DoN task, each associated with a different reward probability (i.e., 20%, 40%, 60%, 80%; Figure 3b). Participants are told the possible outcomes for choosing a risky cue, but must learn its reward probability through experience. Importantly, participants receive feedback about the risky cue on every trial irrespective of their choice; that is, when a participant chooses the certain cue on a given trial, they learn what they would have received had they chosen the risky cue. As such, every trial on the DoN task involves a prediction error (i.e., factual or counterfactual).

Notice that the optimal policy for half the trials on the DoN task is opposite that for other half. Specifically, the risky cue is the reward-maximizing choice for trials involving the 60% and 80% cues, whereas the certain cue is the reward-maximizing choice for trials involving the 20% and 40% cues. This symmetric design is intentional and directly motivated by the analyses in Studies 1a and 1b. Recall how, under unmodelled nonstationary learning, the repeated choice of a cue despite an occasional negative prediction error leads to positively-biased asymmetric learning rates; and similarly how, under identical conditions, the repeated avoidance of a cue despite an occasional position prediction error leads to negatively-biased asymmetric learning rates. In the DoN task, we pit these diametrically-opposed processes against each other in order to effectively neutralize them. Concretely, the positive bias from the repeated choice of the 60%/80% risky cues expected under a nonstationary learning rates model should be counteracted by the negative bias from the repeated avoidance of the 20%/40% risky cues also expected under the same model.

Indeed, this is precisely what we observe in simulation (Figure 3c). Specifically, we simulated choice behavior on the DoN task under a nonstationary learning rates model (using the same empirically-derived parameters as in Study 1a) and fit it with the asymmetric learning rates model. In stark contrast to the 2AFC and RST paradigms, virtually no bias is observed in asymmetric learning rates under each level of nonstationary learning rates (stationary learning: $|\bar{\kappa}| = 0.057$); moderate nonstationarity: $|\bar{\kappa}| = 0.064$; large nonstationarity $|\bar{\kappa}| = 0.073$). Furthermore, recovery of asymmetric learning rates is excellent on the DoN task (Figure S6), even in the presence of unmodeled nonstationary learning (Figure S7). Thus, the DoN task minimizes the threat of nonstationary learning to the unbiased and reliable estimation of asymmetric learning rates.

When we repeat the model recoverability analysis for the DoN task (following the same procedure as in Study 1a and 1b), we find recoverability is markedly improved on the DoN task (Figure 3d). True-positive rates are in excess of 80% for all three models, and is especially improved for the asymmetric learning rates model. This finding likely reflects the fact that the asymmetric learning rates model makes a unique behavioral prediction on the DoN task. In contrast to the nonstationary learning and perseverative choice models, the correlation matrix

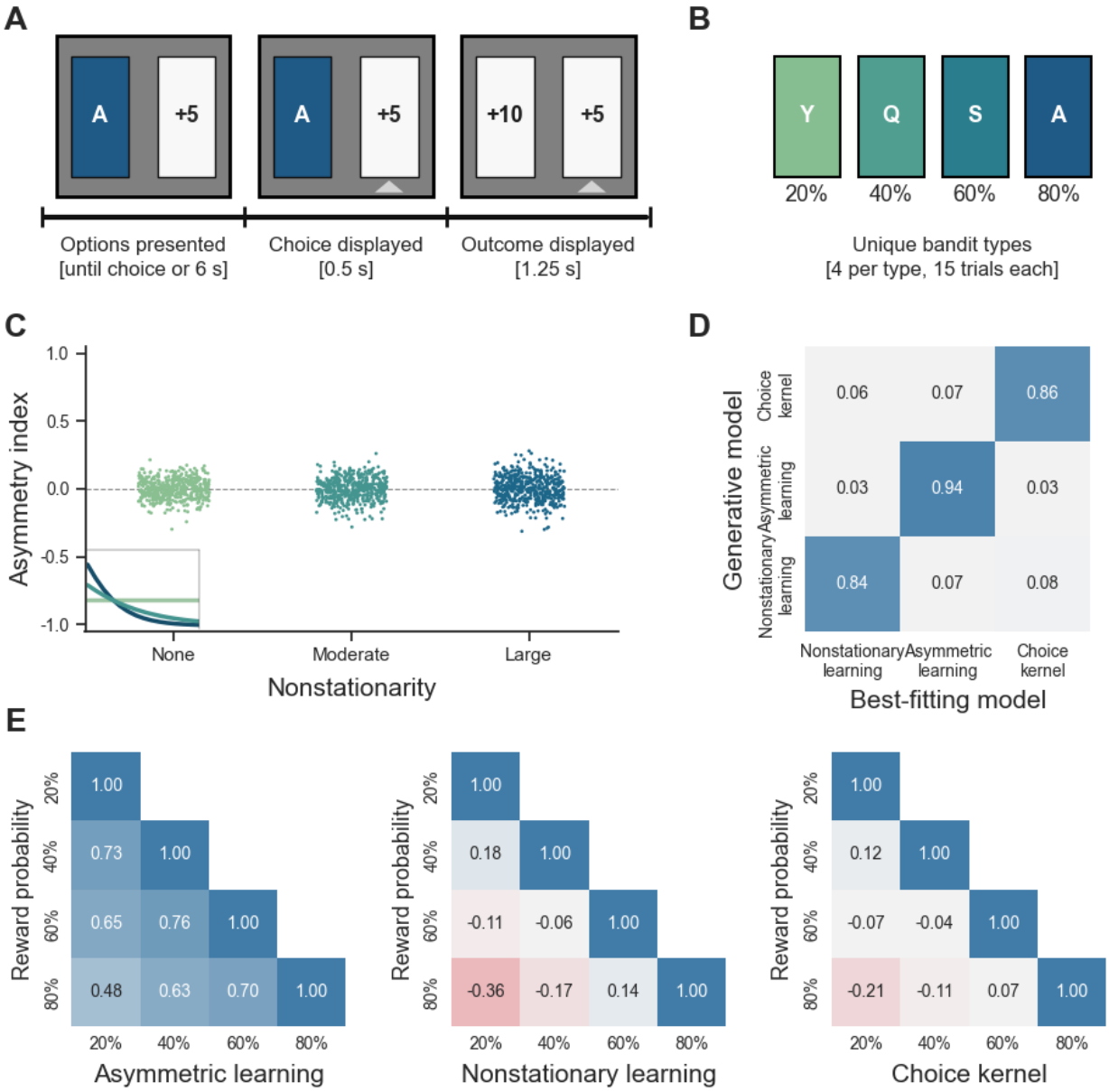

Figure 3: (A) One trial of the Double-or-Nothing task, which always involved a choice between two cues. The risky cue (i.e., a face-down card with a distinguishing symbol on its back) was associated with a large reward (i.e., `+10` points) with fixed probability. The certain cue (i.e., a face-up card) was always associated with a medium reward (i.e., `+5` points). The participant's choice was displayed with a white indicator below the chosen card, and feedback was always provided for the risky cue regardless of the choice. (B) The four risky cue types in the Double-or-Nothing task (i.e., 20%, 40%, 60%, or 80% chance of a large reward). There were four unique instances of each cue type (16 total cues), each presented for 15 trials. (C) Model-predicted asymmetry indices (κ) as a function of unmodeled nonstationary learning. Inset: The latent learning rate curves underlying each condition. (D) Confusion matrix for the model recoverability analysis for the Double-or-Nothing task. Values indicate the proportion of times a data-generating model (y-axis) was also the best-fitting model (x-axis) (E) Pairwise Spearmna's rank correlations of the proportion of risky choices across risky cue types by reinforcement learning model. The asymmetric learning rates model is unique in predicting exclusively positive correlations between the proportion of risky choices across risky cue types.

of the proportion of risky choices across the four cue types is strictly positive under the asymmetric learning rates model 3e). This is because an increased sensitivity to positive (negative) prediction errors increases (decreases) the preference for the risky cue across all reward probability conditions. In sum, not only is the DoN able to dissociate between qualitatively distinct models of choice behavior, it provides a model-agnostic test for whether participants exhibit differential sensitivity to signed prediction errors.

In Study 2, we therefore sought to examine behavior on the DoN task in actual participants. Our primary aim was to determine whether the DoN task produces reliable individual-differences in asymmetric learning rates. We were secondarily interested in whether participants were equally sensitive to factual and counterfactual prediction errors.

2.1 Methods

2.1.1 Participants

147 participants were recruited from Prolific Academic to participate in an online behavioral experiment in November, 2021. Participants were eligible participate if they were at least 18 years old and resided in the United States. The study was approved by the Institutional Review Board of Princeton University (#11968), and all participants provided informed consent. Total study duration was approximately 15-20 minutes. Participants received monetary compensation for their time (rate USD $12/hr), plus an incentive-compatible bonus up to $1.00 based on task performance.

Data from N=27 participants who completed the experiment were excluded prior to analysis (see "Exclusion criteria" below), leaving a final sample of 120 participants. All of these participants were re-invited to complete the DoN task a second time one week later. Once invited, participants were permitted 48 hours to complete the follow-up experiment. A total of N=43 participants completed the retest session. Complete demographic information for the sample is presented in Table S2. Briefly, the majority of participants identified as women (66 women; 53 men; 1 non-binary individual) and were, on average, 28.6 years old (range: 18 – 60 years).

2.1.2 Experimental protocol

In the initial testing session, the study was divided into two parts. After giving consent, participants first completed multiple self-report questionnaires: the 7-item generalized anxiety disorder scale (Spitzer et al., 2006); the 8-item Penn State worry questionnaire (Kertz et al., 2014); the 12-item behavioral inhibition / behavioral activation scales (Pagliaccio et al., 2016); and the general risk question (Dohmen et al., 2011). These measures were included as part of exploratory analyses to measure the associations between mental health, personality, and behavioral risk preferences (Table S3).

After the questionnaires, participants completed the DoN task. In the task, the cues were presented as playing cards. On every trial, participants were presented with two cards: one face-up and one face-down. The face-up card was always worth 5 points, which was presented to participants. The face-down card, which was always worth 10 or 0 points, had a symbol on the back (i.e., a unique animal silhouette). Participants were given up to 6000 ms to make a choice. After their choice, the face-down card was flipped over and revealed what participants earned (if they chose it) or would have earned (if they chose the face-up card). Feedback was presented for 1750 ms. The task was divided into four mini-blocks of 60 trials. In each mini-block, each risky cue type (i.e., 20%, 40%, 60%, 80%) was presented 15 times where the order of trials was randomized across participants under the constraint that no cue could be presented more than twice in a row. In total, participants were exposed to 16 unique risky cues (four per risky cue type) over 240 trials. Participants were offered a break after the first two mini-blocks (120 trials).

Before starting the task, participants were required to review instructions, correctly answer five comprehension questions, and complete several practice trials. Failing to correctly answer all items forced the participant to reread a section of the instructions. After the task, participants completed a standardized feedback form (Hart and Staveland, 1988) where they rated the clarity of the task's instructions, the level of mental effort it required, and frustrating it was to complete. These appraisals are presented in the Supplementary Materials.

The retest session was identical to the initial testing session except for two important differences. First, participants did not complete the self-report questionnaires during the retest session. Second, instead of animal silhouettes, the symbols on the face-down cards were letters from the English alphabet.

The task was programmed in jsPsych (De Leeuw, 2015) and distributed using custom web-application software. The experiment code is available at https://github.com/szorowi1/riskPaper, and the web-software is available at https://github.com/nivlab/nivturk. A playable demo of the task is available at https://nivlab.github.io/jspsych-demos/tasks/mrst/experiment.html.

2.1.3 Exclusion criteria

To ensure data quality, 27 out of 147 recruited participants were excluded prior to analysis for one or more of the following reasons: providing one or more suspicious or incoherent responses to the attention checks embedded in the self-report measures (N=11; Zorowitz et al., n.d.); requiring more than five attempts to complete the comprehension check (N=1); failing to exhibit a sensitivity to the reward probability conditions (defined as a statistically insignificant difference in the proportion of risky choices in the 20% and 80% reward trials; N=14); and failing to exhibit sensitivity to reward outcomes (defined as statistically insignificant win-stay lose-shift choice effect; N=10). This left a total of 120 datasets from the initial test session for analysis. No exclusions were applied to the retest session data.

2.1.4 Reinforcement learning models

To explain participants' choice behavior on the DoN task, we fit a series of five nested reinforcement learning models. We start by explaining the most complex model (M5), which can be reduced all preceding models by fixing certain free parameters. Under M5, the probability that a participant chooses the risky cue, $p(y = 1)$, is defined as:

$$p(y = 1) = \text{logit}^{-1}\left(\beta_1 \cdot [Q(s) - 0.5] + \beta_2 \cdot [C_1(s) - C_2(s)]\right) \tag{8}$$

where $Q(s)$ is the current value estimate of a risky cue; $C_1(s)$ and $C_2(s)$ are binary indicators denoting whether the risky cue was chosen on its last presentation; β_1 is the choice consistency (inverse temperature) parameter; and β_2 is the autoregressive choice parameter. (Here we rescaled rewards to be in the range [0,1], so that value of the certain cue was fixed to 0.5.) In addition, we let the value estimate of a risky cue on first presentation, q_0, be a free parameter.

In turn, the learning rule for M5 was defined as:

$$Q(s) \leftarrow \begin{cases} \eta_{c+} \cdot \delta & \text{if } y = 1 \text{ and } \delta > 0 \\ \eta_{c-} \cdot \delta & \text{if } y = 1 \text{ and } \delta < 0 \\ \eta_{u+} \cdot \delta & \text{if } y = 0 \text{ and } \delta > 0 \\ \eta_{u-} \cdot \delta & \text{if } y = 0 \text{ and } \delta < 0 \end{cases} \tag{9}$$

such that there were four independent learning rates, crossed by prediction error sign (positive, η_{+}; negative, η_{-}) and feedback type (factual, η_c; counterfactual, η_u). This allowed us to test if participants were differentially sensitive to factual and counterfactual feedback. In total, M5 involved seven free parameters per participant (i.e., β_1, β_2, q_0, η_{c+}, η_{c-}, η_{u+}, η_{u-}).

Working backwards, each preceding model involved one simplification to its successor. In Model 4, no distinction was made between factual and counterfactual prediction errors; that is, we assumed $\eta_{c+} = \eta_{u+}$ and $\eta_{c-} = \eta_{u-}$, thereby reducing the model to five free parameters. In Model 3, the choice autocorrelation parameter was fixed to $\beta_2 = 0$, reducing the model to four free parameters. In Model 2, the initial Q-value for the risky cue was fixed to $q_0 = 0.5$. Finally, in Model 1, no distinction was made between positive and negative prediction errors; that is, $\eta_{+} = \eta_{-}$. Model 1 then is simply the canonical Q-learning model.

All models were estimated within a hierarchical Bayesian modeling framework using Hamiltonian Monte Carlo as implemented in Stan (v2.30; Carpenter et al., 2017). For all models, four separate chains with randomized start values each took 6,250 samples from the posterior. The first 5,000 samples from each chain were discarded. Thus, 5,000 post-warmup samples from the joint posterior were retained. The $\hat{R}$ values for all parameters were ≤ 1.01, indicating acceptable convergence between chains, and there were no divergent transitions in any chain. Our choices of prior for each model are detailed in the Supplementary Materials. Briefly, we specified diffuse and uninformative priors in order to avoid biasing parameter estimation.

2.1.5 Goodness-of-fit & model comparison

The fits of the five reinforcement learning models to participants' choice data were assessed using posterior predictive checks. At the group-level, we inspected each model's ability to accurately reproduce the learning curves per reward condition (i.e., the proportion of risky choices participants made per trial and risky cue type). At the participant-level, we inspected each model's ability to accurate predict the overall proportion of risky choices by reward condition. The accuracy of these participant-level predictions were summarized using the root-mean-square error (RMSE) and Spearman correlation between the observed and model-predicted proportion of risky choices.

The fits of the models were compared using Bayesian leave-one out cross-validation (CV-LOO; Vehtari et al., 2017). Following recent best-practice recommendations (Vehtari, n.d.), a credible improvement in model fit was defined as a difference in LOO values four times larger than its corresponding standard error. We also the effective number of parameters (p_{loo}) for each model, which is a measure of model complexity. Conceptually, the effective number of parameters reflects the degree to which model parameters are informed by the data (as opposed to primarily reflecting their priors).

2.1.6 Model parameter reliability calculations

A primary objective of Study 2 was to measure the split-half reliability (within session 1) and test-retest reliability (between sessions 1 & 2) of individual differences across model parameters (especially the asymmetry index, κ). To do so, we used a nested hierarchical modeling approach where parameters were pooled both within- and across-participants (see Rouder and Haaf, 2019). Briefly, for a particular parameter type, separate parameters were estimated per participant and session (in the case of test-retest reliability) or task-half (in the case of split-half reliability). Specifically, the parameters were estimated as follows:

$$\begin{aligned} \theta_{i1} &= \mu_1 + \theta_{ic} - \theta_{id} \\ \theta_{i2} &= \mu_2 + \theta_{ic} + \theta_{id} \end{aligned} \tag{10}$$

where θ_{i1} and θ_{i2} are some parameter (e.g., inverse temperature) for participant i in sessions (or task-halves) 1 and 2, respectively; μ_1 and μ_2 are the group-averaged parameters for sessions (or task-halves) 1 and 2; θ_{ic} is the common effect for participant i (i.e., the parameter component that is stable across sessions or task-halves); and θ_{id} is the difference effect for participant i (i.e., the parameter component that varies across sessions or task-halves). The collection of θ_{ic} parameters across participants represents between-participants variability, whereas the collection of θ_{id} parameters represent within-participants variability. Both θ_{ic} and θ_{id} were assumed to be normally-distributed with zero means and independent, estimated variances. Split-half and test-retest reliability estimates were calculated by taking the Pearson correlation of θ_{i1} and θ_{i2} across task-halves and sessions, respectively.

2.2 Results

2.2.1 Descriptive statistics

The group-average learning curves from the first experiment session are presented in Figure 4a. Participants were, on average, essentially risk-neutral; collapsing across reward conditions, they chose the risky cue on 52.6% of trials – a statistic that is significantly but marginally different than chance (one-sample t-test: $t = 2.868$, $p = 0.005$). Unsurprisingly, participants on average exhibited significant differences in choice behavior by reward condition (one-way ANOVA: $F = 263.054$, $p < 0.001$). The group-average proportion of risky choices on 60%/80% reward trials was significantly greater than that for 20%/40% trials ($\mu_1 = 0.708$, $\mu_2 = 0.345$, $t = 25.336$, $p < 0.001$). These results, however, belie large individual-differences in choice preferences (Figure 4b). Importantly, the proportion of risky choices by reward condition were positively correlated across participants (Figure 4c). All pairwise Pearson correlations were significant at the $\alpha = 0.05$ level (not corrected for multiple comparisons) with the exception of the correlation between 20% and 80% reward conditions ($r = 0.030$, $p = 0.743$). Qualitatively, this pattern of results is most consistent with participants exhibiting differential sensitivity to signed prediction errors than with nonstationary learning or choice perseveration.

An outstanding question is if participants were differentially sensitive to factual and counterfactual feedback. One means of testing this is investigate participants' win-stay lose-shift rates; that is, to determine if participants were differentially likely to choose a given risky cue again after receiving positive factual feedback than to shift from the certain cue after receiving positive counterfactual feedback. Across participants, we did not observe such a difference (paired sample t-test: $t = $ -1.525, $p = 0.132$). A second test is to determine if participants were

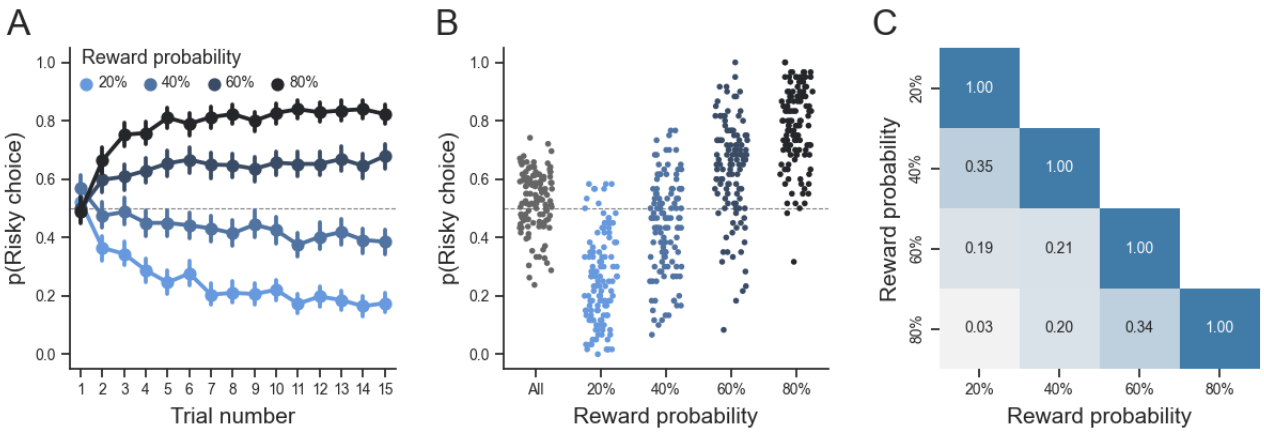

Figure 4: Summary of behavior on the Double-or-Nothing task. (A) The proportion of risky choices across participants as a function of trial number (within a block) and cue type. Error bars indicate 95% bootstrapped confidence intervals. (B) Individual-differences in risky choices. Each dot represents one participant. (C) Pairwise correlations in the proportion of risky choices by cue type across participants. All correlations were significant at the $\alpha = 0.05$ level (not corrected for multiple comparisons) with the exception of the correlation between 20% and 80% cues ($r = 0.030$, $p = 0.743$).

differentially likely to switch to the certain cue after receiving negative factual feedback than to stay with the certain cue after receiving negative counterfactual feedback. Here, we observed a significant difference across participants (paired sample t-test: t = -5.146, $p < 0.001$). The model-agnostic tests were therefore inconclusive as to whether participants treated factual and counterfactual feedback differently. Thus, we turn to the reinforcement learning models to more thoroughly examine this dynamic.

2.2.2 Goodness-of-fit & model comparison

The posterior predictive checks of the model fits to the data are presented in Figure S9 and are summarized in Table 1. The results of the model comparison are also presented in Table 1. There are several noteworthy features of the model comparison results. First, a model with asymmetric learning rates (M2) vastly outperforms a model without (M1). Second, the addition of initial Q-value (q_0) and choice autocorrelation (β_2) parameters further improves model fit. Finally, the overall best-fitting model (M5) involves four independent learning rates by prediction error sign (positive, negative) and feedback type (factual, counterfactual).

This last result warrants further discussion. Though the model comparison shows the inclusion of separate learning rates by feedback type credibly improves model fit, the magnitude of this improvement is marginal at best. Despite the addition of two additional parameters per participant, the smallest improvements in the posterior predictive checks (as quantified by the RMSE and correlation between observed and model-predicted proportion of risky choices per participant and reward condition) are observed between M4 and M5. Similarly, the effective number of parameters increases linearly with each additional parameter (from M1 to M4) until the inclusion of independent factual and counterfactual learning rates, indicating the addition of these parameters are comparatively less informed by the data. As evidence of this, the difference in the group-level factual and counterfactual learning rates (i.e., collapsing across

Model	Parameters	RMSE	r	p_{loo}	LOO-CV	Δ LOO (SE)
1	β_1, η	0.127	0.872	18.8	-37184.1	3545.3 (54.0)
2	β_1, η_+, η_-	0.091	0.936	36.4	-37998.7	2730.7 (48.3)
3	β_1, q_0, η_+, η_-	0.080	0.950	55.8	-38778.4	1951.0 (41.3)
4	β_1, β_2, q_0, η_+, η_-	0.061	0.974	72.0	-40062.7	666.8 (23.1)
5	β_1, β_2, q_0, η_{c+}, η_{c-}, η_{u+}, η_{u-}	0.052	0.981	78.7	-40729.4	-

Table 1: Summary of model fit and comparison for the reinforcement learning models. Notes: β_1 = inverse temperature; β_2 = choice autocorrelation; q_0 = initial Q-value; η_+ = positive learning rate; η_- = negative learning rate; η_c = factual learning rate; η_u = counterfactual learning rate; RMSE = root-mean-square error; r = Pearson correlation; p_{loo} = effective number of parameters; LOO-CV = leave-one-out cross-validation. LOO-CV values are presented in deviance scale, such that smaller numbers indicate better fit.

positive and negative learning rates) is not significantly different than zero ($\mu_{\eta_c} = 0.188$, $\mu_{\eta_u} = 0.205$, $p = 0.098$). At the participant-level, only two of 120 participants exhibited significantly learning rates for factual and counterfactual feedback. In sum, though there appears to be some evidence for differential sensitivity to prediction errors by feedback type, it is weak at best. Therefore, in all following analyses, we proceed by investigating Model 4. (We note, however, this choice does not affect our conclusions; see Table S4).

2.2.3 Group- & individual-level parameter estimates

The group-level parameters for M4 are summarized in Table 2. The distributions of parameters across participants, and their pairwise correlations, are presented in Figure S10. It is worthwhile to note that, at the group-level, the choice autocorrelation parameter (μ_{β_2}) was significantly greater than zero (one-sample t-test: $t = 19.976$, $p < 0.001$), indicating some tendency towards perseverative choice across participants. However, this effect was on average an order of magnitude smaller than choice consistency, suggesting that participants' choices were governed more strongly by the history of rewards than by their past choices.

Most important to present purposes, the group-average asymmetry index (μ_κ) was marginally but significantly greater than zero (one-sample t-test: $t = 3.327$, $p < 0.001$), in line with what was observed in the choice data. There were, however, considerable individual-differences in the asymmetry index across participants. It should be noted that the variance in these estimates is larger than what we observed when fitting the asymmetric learning rates model to data generated under nonstationary learning, providing additional evidence for differential sensitivity to prediction errors in this sample. Crucially, participants' asymmetry indices are significantly correlated with the their corresponding proportion of risky choices across trials ($\rho = 0.504$, $p < 0.001$; Figure 5a). The strength of this correlation further improves when looking at risky choices only in the 40%/60% reward conditions (i.e., where there is the greatest expected variation in prediction errors; $r = 0.569$, $p < 0.001$; Figure 5b).

Parameter	Mean (SD)	Split-half reliability	Test-retest reliability
β_1	7.71 (3.35)	0.77 (0.69 – 0.84)	0.75 (0.60 – 0.86)
β_2	0.68 (0.37)	0.90 (0.86 – 0.93)	0.92 (0.86 – 0.96)
q_0	0.52 (0.10)	0.99 (0.99 – 0.99)	0.95 (0.91 – 0.98)
η_+	0.15 (0.03)	0.98 (0.98 – 0.99)	0.96 (0.92 – 0.97)
η_-	0.13 (0.01)	0.81 (0.72 – 0.87)	0.46 (0.15 – 0.71)
κ	0.04 (0.14)	0.96 (0.94 – 0.98)	0.72 (0.57 – 0.83)

Table 2: Summary and reliability of the reinforcement learning model parameters (M4) across participants. For the reliability estimates, values in parentheses denote the 95% confidence intervals.

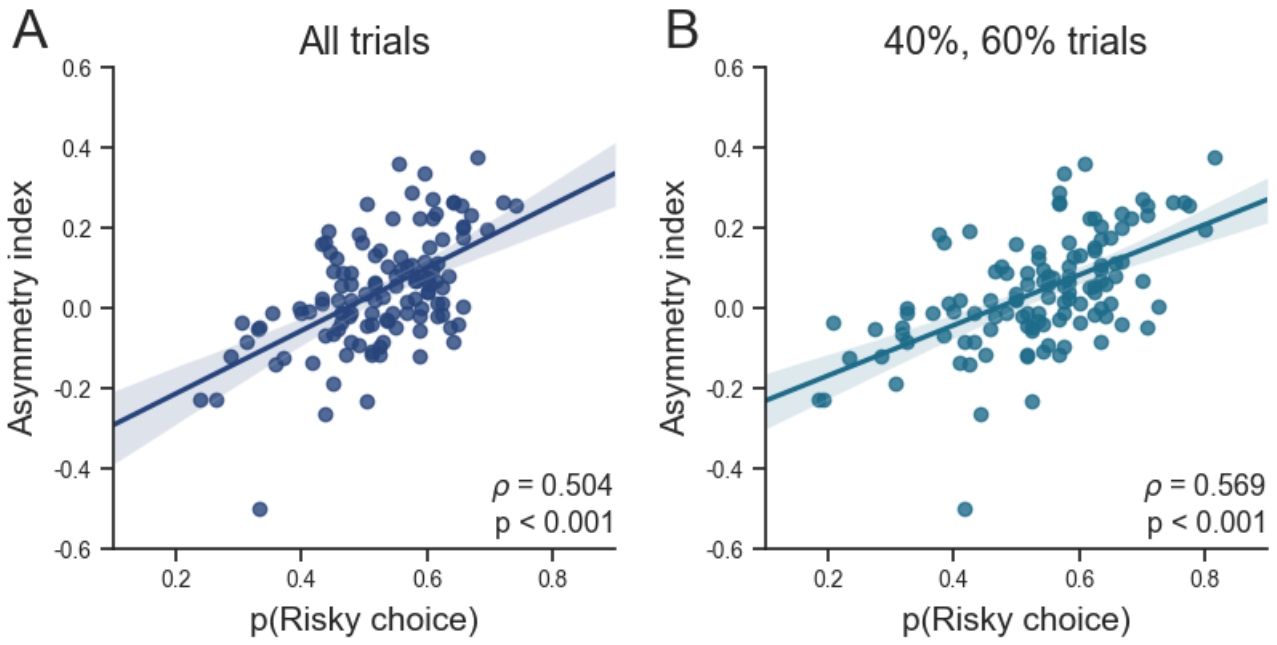

Figure 5: Pearson correlations between the proportion of risky choices and asymmetry indices across participants. (A) Proportions calculated across all trials. (B) Proportions calculated from only 40%/60% reward probability trials.

2.2.4 Parameter stability & reliability

Within the first session, the group-level parameter estimates were mostly stable across task-halves. Across halves, the changes in the inverse temperature ($\Delta\mu_{\beta_1}$ = -0.115, 95% HDI = [-1.138, 0.847]), choice autocorrelation ($\Delta\mu_{\beta_2}$ = 0.051, 95% HDI = [-0.034, 0.137]), positive learning rate ($\Delta\mu_{\eta_+}$ = 0.001, 95% HDI = [-0.018, 0.018]), and negative learning rate ($\Delta\mu_{\eta_-}$ = 0.006, 95% HDI = [-0.010, 0.022]) were not credibly different than zero. Only the initial Q-value showed a small but credible decrease from the first to the second half of the task ($\Delta\mu_{q_0}$ = -0.022, 95% HDI = [-0.041, -0.005]). A similar pattern of results was observed across sessions. Between sessions, there was no credible difference in estimates for the inverse temperature ($\Delta\mu_{\beta_1}$ = -1.378, 95% HDI = [-0.229, 3.145]), choice autocorrelation ($\Delta\mu_{\beta_2}$ = -0.069, 95% HDI = [-0.175, 0.034]), initial Q-value ($\Delta\mu_{q_0}$ = -0.022, 95% HDI = [-0.050, 0.009]), positive learning rate ($\Delta\mu_{\eta_+}$ = -0.008, 95% HDI = [-0.034, 0.018]), or negative learning rate ($\Delta\mu_{\eta_-}$ = -0.011, 95% HDI = [-0.034, 0.012]). Thus, there was little evidence for systematic shifts in choice behavior on the DoN task (e.g., due to practice effects) both within- and across-sessions.

Finally, the split-half and test-retest reliability estimates for each model parameter are presented in Table 2. As is expected, the split-half reliability estimates are, on average, larger than their corresponding test-retest reliability estimates. Though we do not advocate for the use of arbitrary cutoffs for reliability, we note that all estimates of split-half reliability are in excess of $\rho = 0.7$, which is sometimes suggested as the cutoff for "good" reliability for use in individual-differences research Parsons et al., 2019. Importantly, this extends to the asymmetry index (κ). The same pattern of results was observed for the test-retest reliability estimates, with the exception of the negative learning rate, η_-. The discrepancy in reliability between the

positive and negative learning rates likely reflects the smaller between-participants variability for the latter. Regardless, the test-retest reliability estimate for asymmetry index is still good by conventional standards. In sum, reinforcement learning model parameters estimated from choice behavior on the DoN task in this current sample exhibit was both stable, within and across sessions, and exhbiited acceptable reliability for individual-differences research.

3 General Discussion

References

Arkadir, D., Radulescu, A., Raymond, D., Lubarr, N., Bressman, S. B., Mazzoni, P., & Niv, Y. (2016). Dyt1 dystonia increases risk taking in humans. *Elife*, *5*, e14155.

Aylward, J., Valton, V., Ahn, W.-Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature human behaviour*, *3*(10), 1116–1123.

Bennett, D., & Niv, Y. (2020). Opening Burton's clock: Psychiatric insights from computational cognitive models. In *The cognitive neurosciences* (pp. 439–250). The MIT Press.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32.

Cavanagh, J. F., Bismark, A. W., Frank, M. J., & Allen, J. J. (2019). Multiple dissociations between comorbid depression and anxiety on reward and punishment processing: Evidence from computationally informed EEG. *Computational Psychiatry (Cambridge, Mass.)*, *3*, 1.

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Cockburn, J., & Holroyd, C. B. (2010). Focus on the positive: Computational simulations implicate asymmetrical reward prediction error signals in childhood attention-deficit/hyperactivity disorder. *Brain research*, *1365*, 18–34.

Craig, S., Lewandowsky, S., & Little, D. R. (2011). Error discounting in probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 673.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association*, *9*(3), 522–550.

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(3), 490–502.

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: Impulsivity, deep brain stimulation, and medication in Parkinsonism. *science*, *318*(5854), 1309–1312.

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *Journal of Neuroscience*, *38*(36), 7901–7911.

Harada, T. (2020). Learning from success or failure?–positivity biases revisited. *Frontiers in Psychology*, *11*, 1627.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183). Elsevier.

Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, *87*, 31–45.

Kertz, S. J., Lee, J., & Björgvinsson, T. (2014). Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychological assessment*, *26*(4), 1146.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9.

Litman, L., Rosenzweig, C., & Moss, A. (2020). New solutions dramatically improve research data quality on MTurk.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, *15*(7), 1040–1046.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.

Ossola, P., Garrett, N., Sharot, T., & Marchesi, C. (2020). Belief updating in bipolar disorder predicts time of recurrence. *Elife*, *9*.

Pagliaccio, D., Luking, K. R., Anokhin, A. P., Gotlib, I. H., Hayden, E. P., Olino, T. M., Peng, C.-Z., Hajcak, G., & Barch, D. M. (2016). Revising the BIS/BAS Scale to study development: Measurement invariance and normative effects of age and sex from childhood through adulthood. *Psychological assessment*, *28*(4), 429.

Palminteri, S. (2022). Choice-confirmation bias and gradual perseveration in human reinforcement learning. *Behavioral Neuroscience*.

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in cognitive sciences*.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395.

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532.

Radulescu, A., Holmes, K., & Niv, Y. (2020). On the convergent validity of risk sensitivity measures.

Rosenbaum, G. M., Grassie, H. L., & Hartley, C. A. (2022). Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. *Elife*, *11*, e64620.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review*, *26*(2), 452–467.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of internal medicine*, *166*(10), 1092–1097.

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific reports*, *11*(1), 1–13.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vehtari, A. (n.d.). Cross-validation FAQ [Accessed: 2023-2-2].

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Zorowitz, S., Niv, Y., & Bennett, D. (n.d.). Inattentive responding can induce spurious associations between task behavior and symptom measures.

Supplementary materials

Nonstationary learning model provides acceptable fit to choices from the archival 2AFC datasets

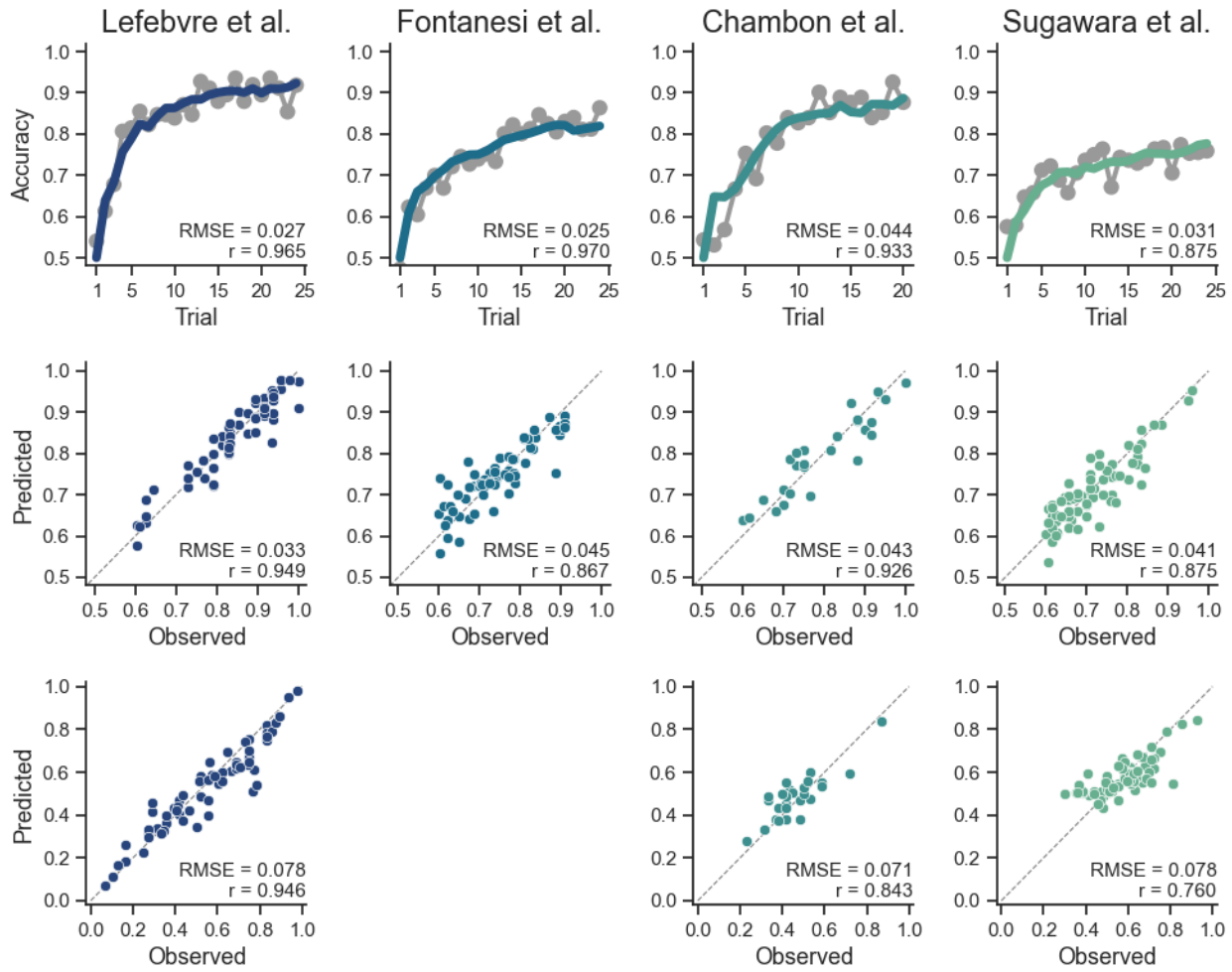

Figure S1: Posterior predictive checks of the fit of the nonstationary learning model to the 2-alternative forced choice (2AFC) datasets. *Top row:* Observed (grey) and predicted (colored) group-average learning curves in the 2AFC blocks with unequal reward probabilities (e.g. 75%/25%). Accuracy denotes the proportion of choices of the higher value arm. *Middle row:* Observed (x-axis) and predicted (y-axis) proportion of choices of the higher value arm in the 2AFC blocks with unequal reward probabilities. Each point represents one participant. *Bottom row:* Observed (x-axis) and predicted (y-axis) proportion of choices of one arm in the 2AFC blocks with equal reward probabilities (e.g. 50%/50%). Each point represents one participant. *Notes:* Degree of fix quantified using the root-mean-square error (RMSE) and Pearson correlation (r) between the observed and predicted values.

Nonstationary learning model provides acceptable fit to choices from the archival RST datasets

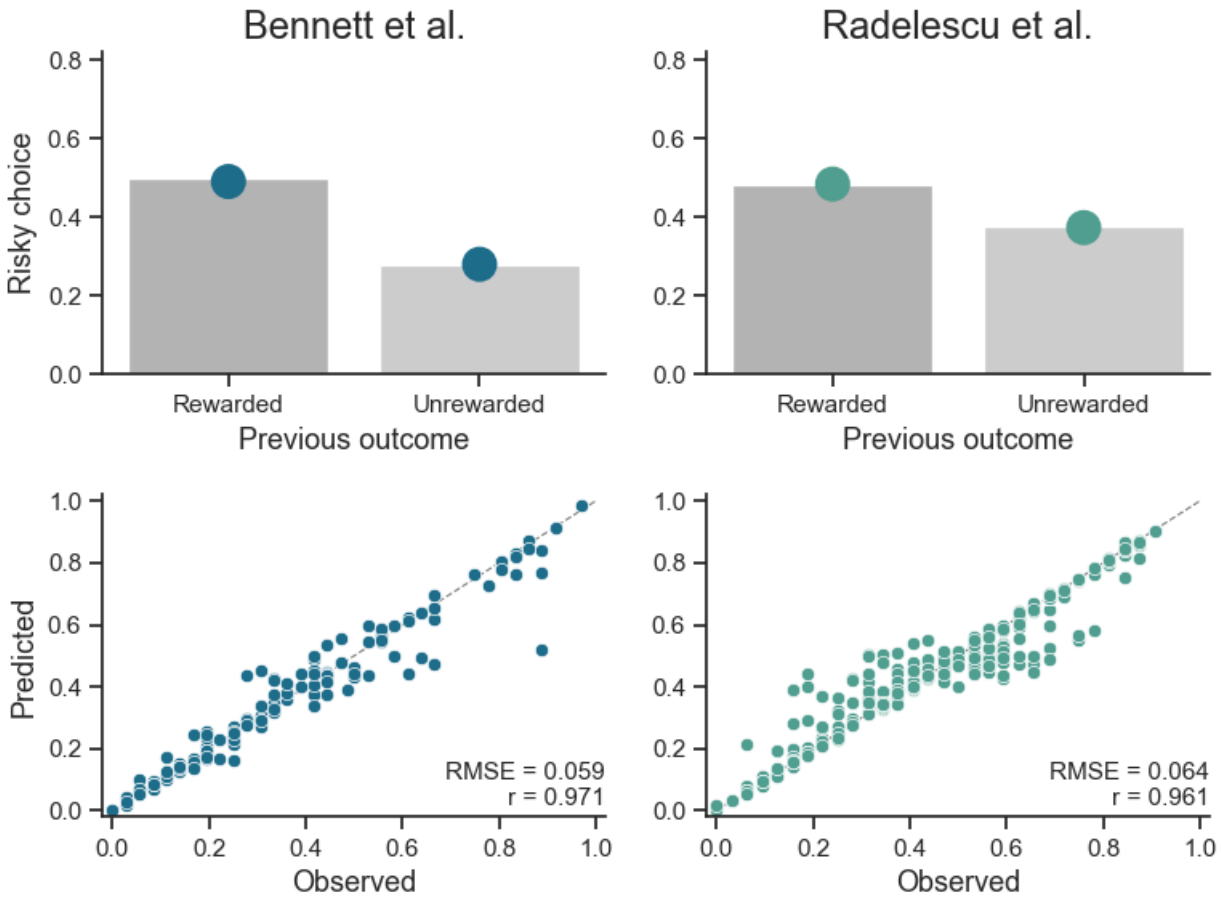

Figure S2: Posterior predictive checks of the fit of the nonstationary learning model to the risk sensitivity task (RST) datasets. *Top row:* Observed (grey) and predicted (colored) group-average risky choices on risky vs. certain trials following a previously rewarded or unrewarded choice of the risky bandit. *Bottom row:* Observed (x-axis) and predicted (y-axis) proportion of risky choices on risky vs. certain trials. Each point represents one participant. *Notes:* Degree of fix quantified using the root-mean-square error (RMSE) and Pearson correlation (r) between the observed and predicted values.

Parameter bias in 2AFC paradigm from unmodeled nonstationarity moderated by number of trials and lower asymptote of learning rate

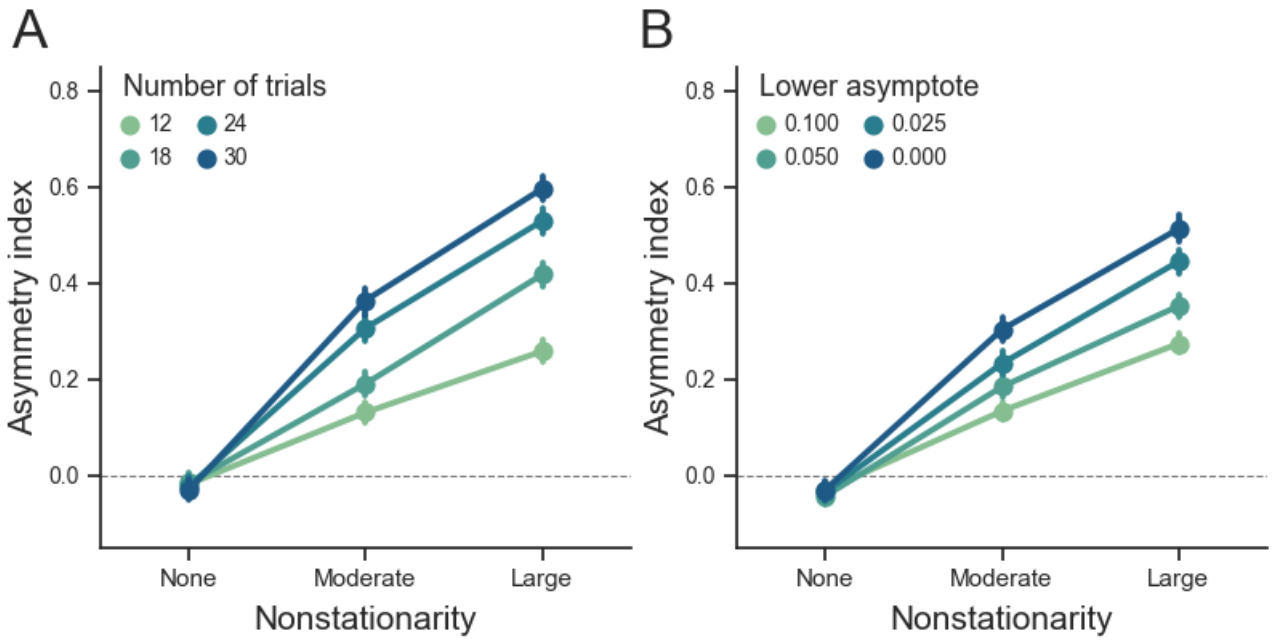

Figure S3: Moderators of parameter bias in the 2AFC paradigm from unmodeled nonstationarity. (A) As the number of trials in a block increases, unmodeled nonstationarity produces increasingly positively-biased asymmetric learning rates. Each simulation condition involved approximately the same number of total trials (i.e. 16 blocks of 12 trials, 10 blocks of 18 trials, 8 blocks of 24 trials, 6 blocks of 30 trials). (B) As the lower asymptote of a nonstationary learning curve approaches zero, unmodeled nonstationarity produces increasingly positively-biased asymmetric learning rates. *Methods:* Choice data were generated under the nonstationary learning rates model with the following parameters: stationary: β = 7.5, a_0 = -1.821, a_1 = 0.000; moderate nonstationarity: β = 7.5, a_0 = -1.245, a_1 = -0.179; large nonstationarity: β = 7.5, a_0 = -0.677, a_1 = -0.353. The data were fit with asymmetric learning rates model using *maximum a posteriori* estimation using Stan (v2.30). *Notes:* Error bars denote 95% bootstrapped confidence intervals.

Unmodeled nonstationarity masquerades as perseveration in the 2AFC and RST paradigms

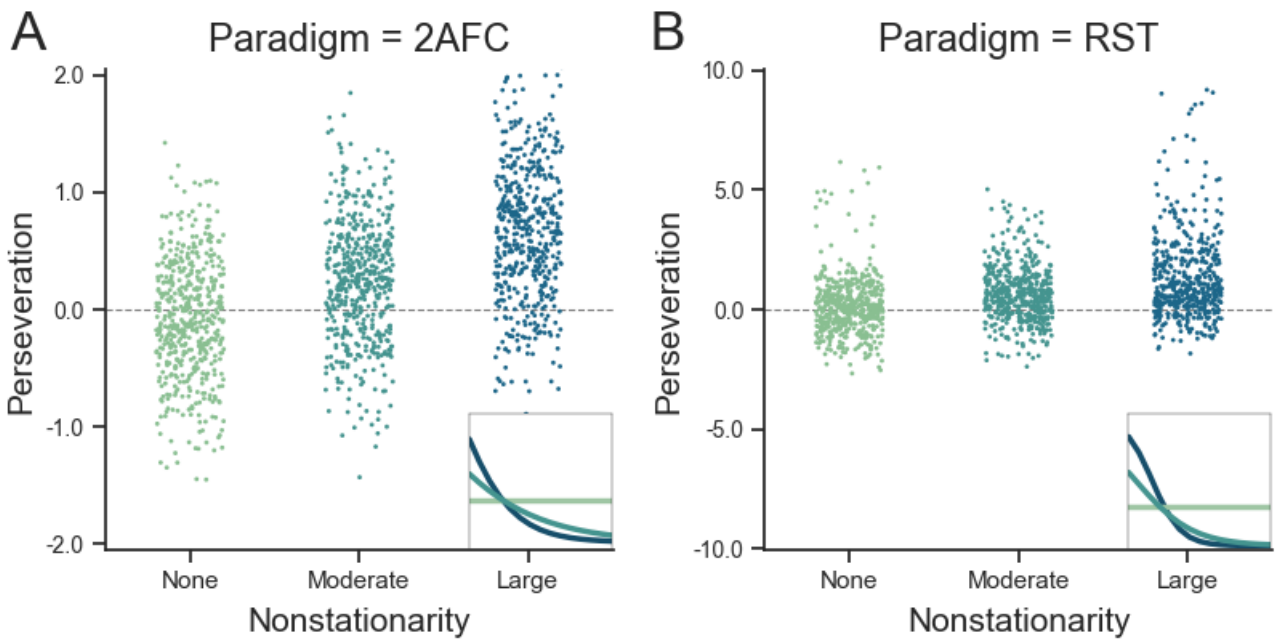

Figure S4: Unmodeled nonstationarity can masquerade as perseveration in the 2-alternative forced choice (2AFC; Panel A) and risk sensitvity task (RST; Panel B) paradigms. As latent nonstationarity increases (i.e. increasing steepness of learning rate decay), estimates of the impulse choice kernel parameter (β_2) are increasingly positively-biased. *Methods:* Choice data were generated under the nonstationary learning rates model using parameters based on empirical priors estimated from archival data. The data were fit with impulse choice kernel model using *maximum a posteriori* estimation using Stan (v2.30).

The graded (but not impulse) perseverative choice model can confuse reward-guided choice for choice perseveration

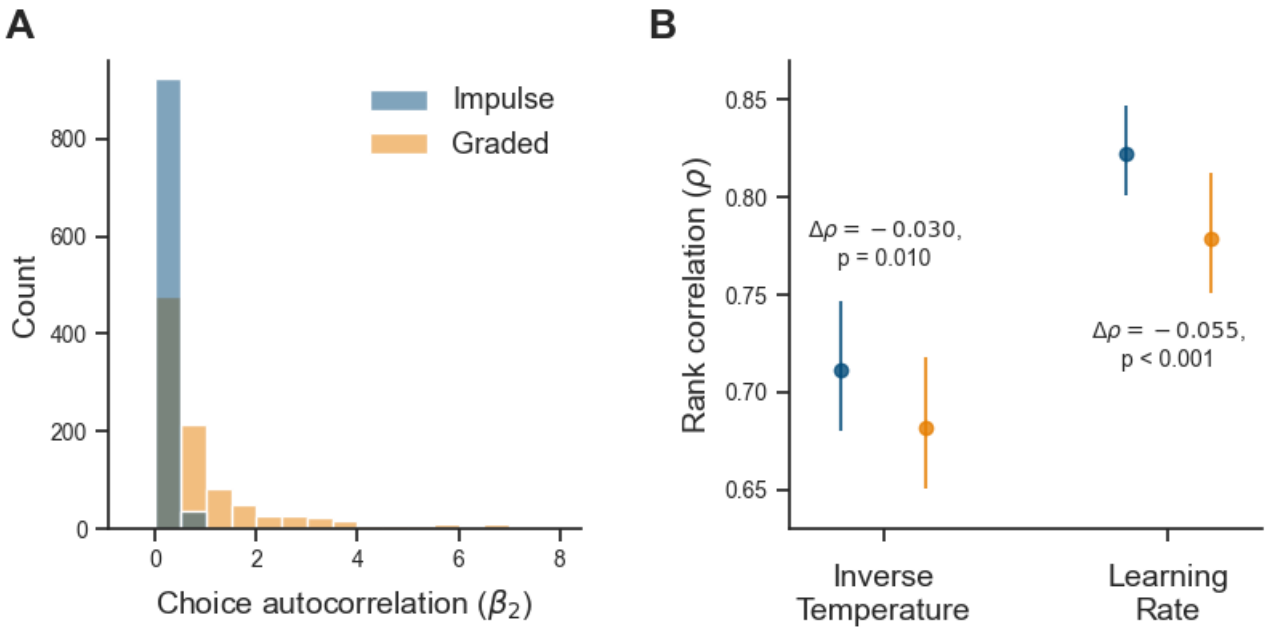

Figure S5: (A) The distribution of choice autocorrelation parameters (β_2) for the impulse (one-trial-back) and graded perseverative choice models when fitted to choice data simulated under the canonical Q-learning model. Even in the absence of choice perseveration, the graded (but not impulse) perseverative choice model frequently indicates moderate-to-large perseveration. (B) Spearman's rank correlation between true and recovered inverse temperature (β_1) and learning rate (η) parameters when fitting the impulse and graded choice perseveration models to choice data simulated under the canonical Q-learning model. Parameter recovery is worse under the graded perseverative choice model. *Methods:* Choice data on the 2AFC task were simulated for under the canonical Q-learning model. The 2AFC task statistics were identical to those used in Study 1a. We explored 961 unique combinations of the inverse temperature (i.e., 31 linearly-spaced values between 5 and 15) and learning rate (i.e., 31 linearly-spaced values between 0.05 and 0.40). The impulse (one-trial-back) and graded perseverative choice models were fitted to each dataset via *maximum a posteriori* estimation using the L-BFGS algorithm as implemented in cmdstan (v2.30).

The Double-or-Nothing task exhibits satisfactory parameter recovery

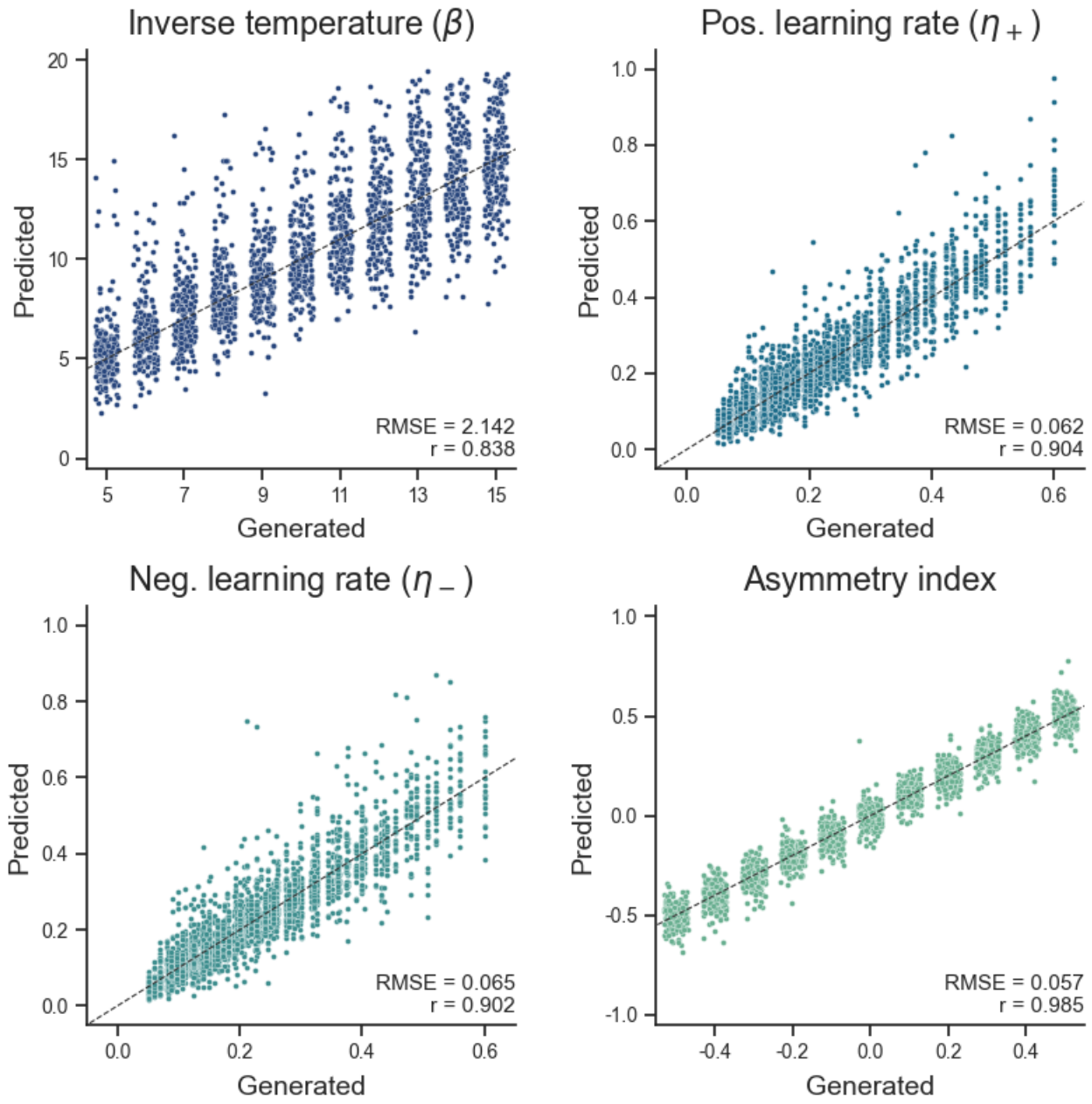

Figure S6: Parameter recovery for the parameters of the asymmetric learning rates model on the Double-or-Nothing task. *Notes:* The true data-generating and recovered parameters are plotted on the x- and y-axes, respectively. Each point represents one simulation. The degree of recovery was quantified using the root-mean-square error (RMSE) and Spearman correlation (r) between the true and recovered values.

Unmodeled nonstationarity minimally impacts the estimation of asymmetric learning rates on the Double-or-Nothing task

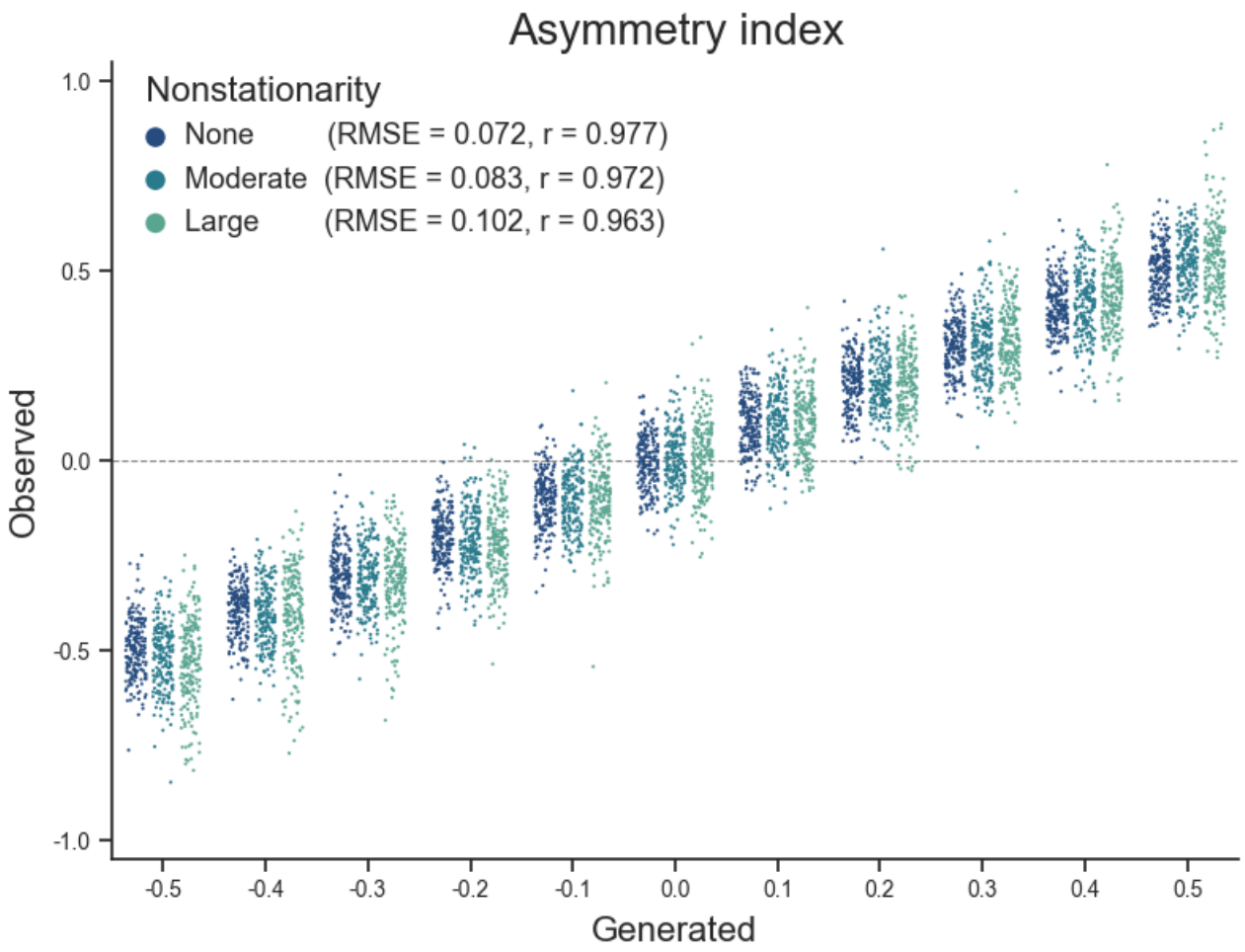

Figure S7: Recovery of the asymmetry index for the asymmetric learning rates model on the Double-or-Nothing task under unmodeled nonstationarity. Increasing nonstationarity (i.e. increasing steepness of learning rate decay) marginally increases the variability in recovered parameters. *Notes:* The true data-generating and recovered parameters are plotted on the x- and y-axes, respectively. Each point represents one simulation. The degree of recovery was quantified using the root-mean-square error (RMSE) and Spearman correlation (r) between the true and recovered values.

The Double-or-Nothing task is better able to discriminate between alternative reinforcement learning models

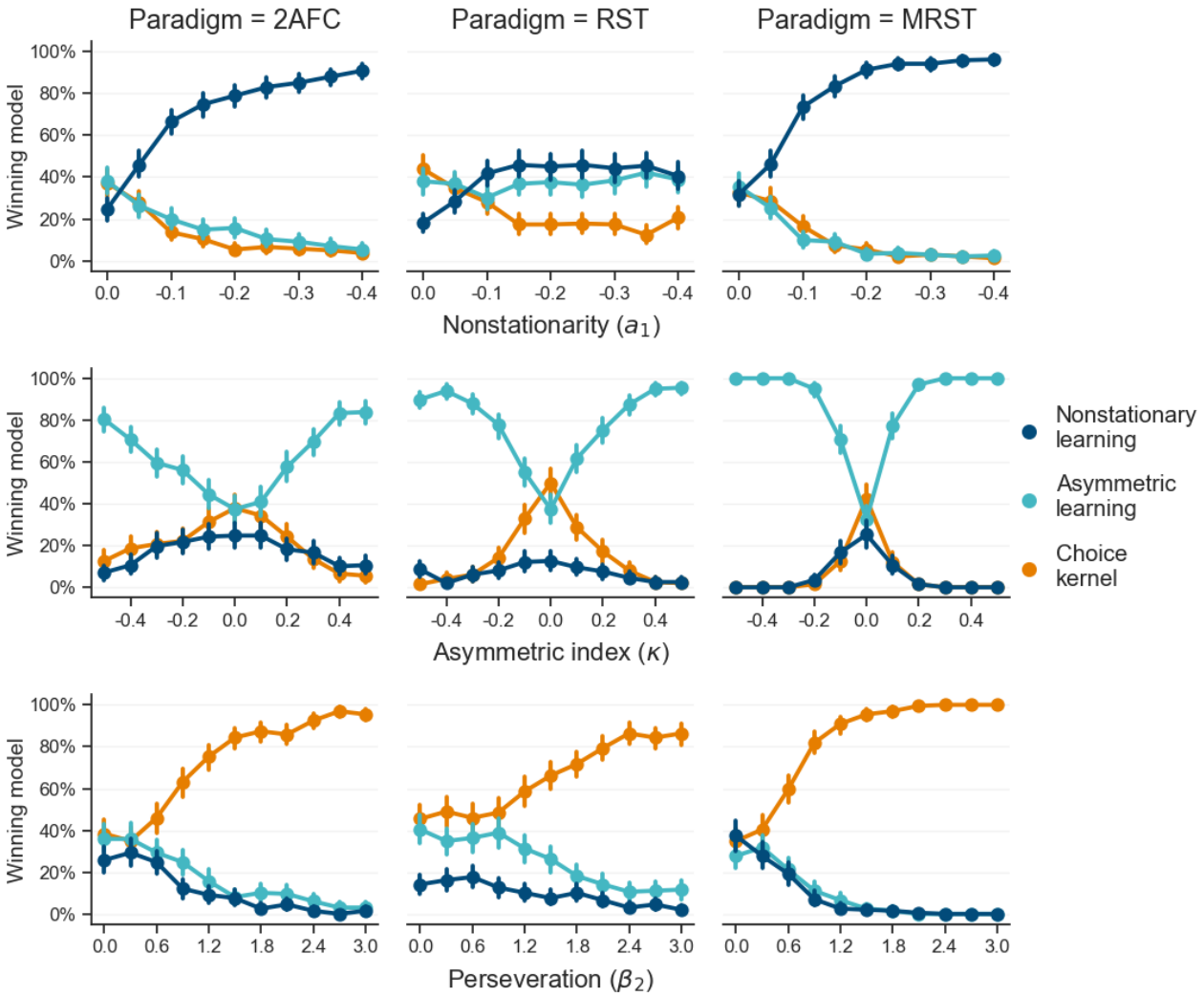

Figure S8: Top row: The percentage of simulations in which the nonstationary learning rates model is correctly identified as the best-fitting model (y-axis) as a function of nonstationarity (a_1; x-axis) and task paradigm. *Middle row:* The percentage of simulations in which the asymmetric learning rates model is correctly identified as the best-fitting model (y-axis) as a function of learning rate asymmetry (κ; x-axis) and task paradigm. *Bottom row:* The percentage of simulations in which the impulse choice kernel model is correctly identified as the best-fitting model (β_2; y-axis) as a function of perseveration (x-axis) and task paradigm. *Notes:* When each of the pivotal parameters are equal to zero, the competing models are equivalent and are thus expected to perform equally well (i.e. 33%).

Posterior predictive checks for the the reinforcement learning models

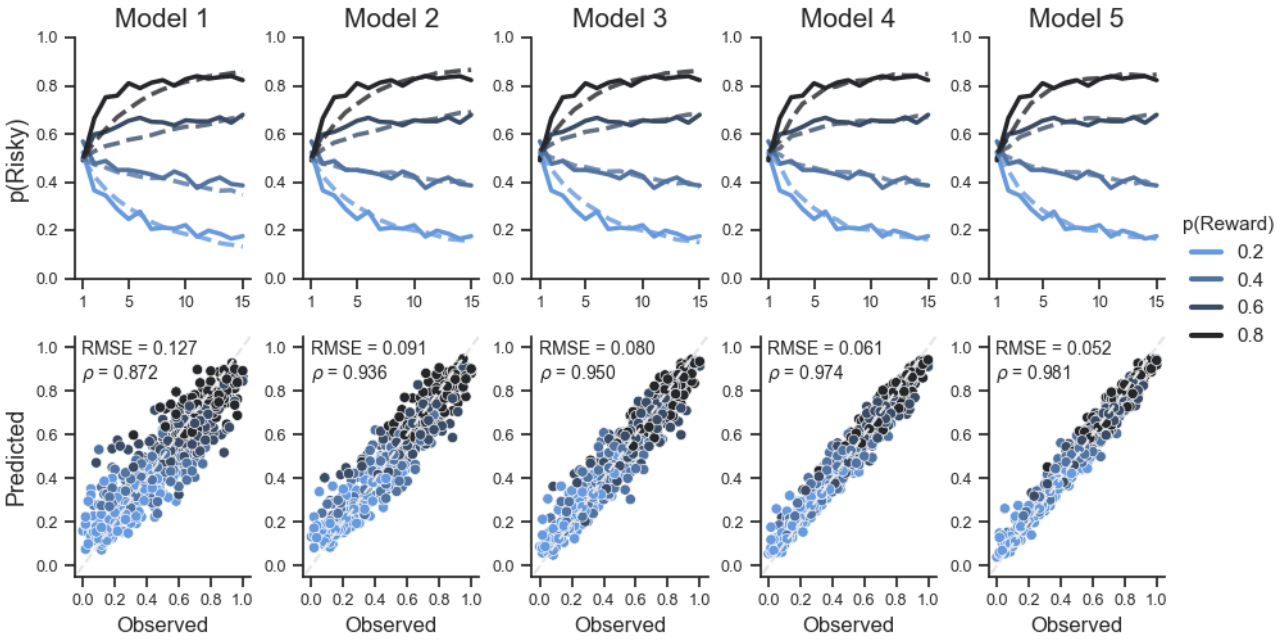

Figure S9: Posterior predictive checks for the reinforcement learning models fitted to the Session 1 choice data. *Top row:* Observed (solid) and model-predicted (dotted) group-averaged learning curves for each bandit type. *Bottom row:* Observed (x-axis) and model-predicted (y-axis) proportions of risky choice per participant and bandit type. Notes: RMSE = root-mean-squared error; p = Spearman's rank correlation.

The distributions and correlations of parameter estimates for Model 4

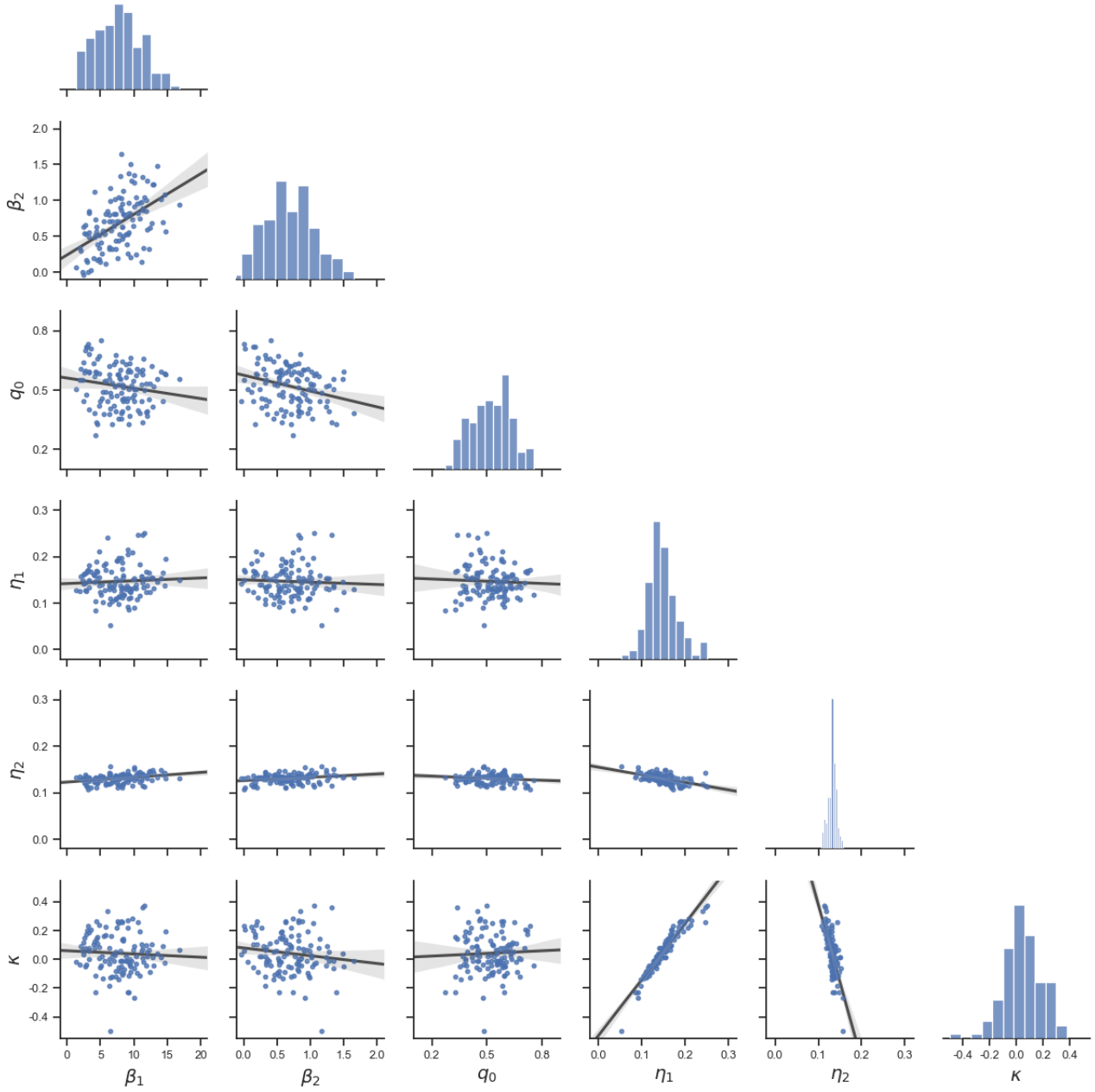

Figure S10: The distributions (diagonal) and correlations (lower-diagonal) of model parameter estimates for the risk-sensitive temporal difference learning model (M4) as fit to the Session 1 data. Notes: β_1 = choice sensitivity (inverse temperature); β_2 = choice autocorrelation (impulse choice kernel); q_0 = initial Q-value; η_+ = positive learning rate; η_- = negative learning rate; κ = asymmetry index.

Range of parameters used for the model recoverability analyses

Model	Parameter	Start	Stop	Step	Count
Asymmetric learning	Inverse temperature (β_1)	5.0	15.0	1.0	11
	Average learning rate (η)	0.1	0.4	0.0375	9
	Asymmetry index (κ)	-0.5	0.5	0.1	11
Nonstationary learning	Inverse temperature (β_1)	5.0	15.0	1.0	11
	Learning rate intercept (a_0)	-2.0	0.0	0.2	11
	Learning rate slope (a_1)	-0.4	0.0	0.05	9
Impulse choice kernel	Inverse temperature (β_1)	5.0	15.0	1.0	11
	Choice autocorrelation (β_2)	0.0	3.0	0.3	11
	Learning rate (η)	0.1	0.4	0.0375	9

Table S1: The range of parameters used to simulate choice data on each task paradigm for each reinforcement learning model. For each parameter, a range of linearly spaced values was selected (e.g., from 5.0 to 15.0 in steps of 1.0). Choice data was then simulated, twice, for each unique combination of parameters (2 x 9 x 11 x 11 = 2178 total).

Participant demographics

Characteristic	Frequency (N)	Percentage (%)
Gender		
Women	66	55.0
Men	53	44.2
Nonbinary	1	0.8
Age		
18 – 29	76	63.3
30 – 39	31	25.8
40 – 49	9	7.5
50 – 59	3	2.5
60 and older	1	0.8
Race & Ethnicity		
White	94	72.3
Hispanic or Latino	16	13.3
Asian	12	9.2
Black or African American	11	8.5
American Indian or Alaska Native	5	3.8
Native Hawaiian or other Pacific Islander	2	1.5
Rather not say	6	4.6
Education		
Less than high school	2	1.7
High school	12	10.0
Some college	43	35.8
Associate degree	8	6.7
Bachelor degree	42	35.0
Master degree or higher	13	10.8

Table S2: Demographic characteristics of the participants in Experiment 2. Note: Participants were able to select more than one ethnic and racial identity. Therefore, the participant counts and percentages in this section sum up to more than N=120 and 100%, respectively.

Spearman correlations between model parameters (M4) & self-report questionnaire scores

Variable	GAD-7	PSWQ	BIS	BAS	GRQ	β_1	β_2	q_0	η_1	η_2	κ
GAD-7	-										
PSWQ	**0.77**	-									
BIS	**0.47**	**0.62**	-								
BAS	**-0.25**	-0.16	-0.12	-							
GRQ	-0.16	**-0.26**	**-0.30**	**0.34**	-						
β_1	-0.02	-0.11	-0.00	**-0.31**	0.05	-					
β_2	-0.02	-0.09	-0.09	**-0.20**	-0.01	**0.52**	-				
q_0	-0.12	0.01	-0.00	0.18	0.08	-0.17	**-0.26**	-			
η_1	0.04	-0.00	0.14	0.04	0.05	0.00	-0.08	-0.02	-		
η_2	-0.02	-0.05	0.06	-0.10	-0.05	**0.38**	**0.24**	-0.12	**-0.58**	-	
κ	0.03	0.00	0.11	0.05	0.05	-0.10	-0.12	0.02	**0.98**	**-0.71**	-

Table S3: Spearman's rank correlation between the model parameters and self-report questionnaire scores. Notes: GAD-7 = generalized anxiety disorder scale; PSWQ = Penn State worry questionnaire; BIS = behavioral inhibition scale; BAS = behavioral activation scale; GRQ = general risk question. Bolded cells denote $p < 0.05$. Correlations are exploratory and not corrected for multiple comparisons.

Summary of the parameter estimates for Model 5

Parameter	Mean (SD)	Split-half reliability	Test-retest reliability
β_1	5.92 (2.44)	0.88 (0.83 – 0.92)	0.90 (0.84 – 0.94)
β_2	0.23 (0.17)	0.97 (0.96 – 0.98)	0.98 (0.96 – 0.99)
q_0	0.52 (0.14)	0.99 (0.99 – 0.99)	0.97 (0.95 – 0.99)
η_{c+}	0.33 (0.01)	0.29 (0.08 – 0.50)	0.70 (0.50 – 0.83)
η_{c-}	0.10 (0.03)	0.98 (0.97 – 0.99)	0.67 (0.42 – 0.84)
η_{u+}	0.13 (0.07)	0.99 (0.99 – 0.99)	0.98 (0.97 – 0.99)
η_{u-}	0.33 (0.04)	0.96 (0.95 – 0.97)	0.85 (0.75 – 0.92)
κ	0.03 (0.11)	0.95 (0.92 – 0.97)	0.85 (0.74 – 0.92)

Table S4: Summary of the distribution and reliability of estimated parameters for Model 5. The parenthetical values next to the reliability estimates are the corresponding 95% confidence intervals.

Double-or-Nothing task timing and appraisal

On average, participants completed the Double-or-Nothing task in approximately 10 minutes (session 1: mean = 10.7 min, range = 9.3 – 15.2; session 2: mean = 10.4 min, range = 9.3 – 11.9) and the instructions in 3 – 4 minutes (session 1: mean = 3.5 min; range = 1.7 – 11.4; session 2: mean = 3.9 min, range = 1.4 – 26.4). We note that, as opposed to the task, the instructions were self-paced, meaning participants were able to take breaks if they so choose. Participants were able to complete the entire experiment, on average, in 14 minutes.

On a scale of 1 ("Not at all") to 7 ("Very much"), the majority of participants rated the task instructions as very clear (mean = 6.4, sd = 1.9). On the same scale, participants appraised the task as requiring moderate mental demand (mean = 3.7, sd = 2.8) and to be moderately frustrating (mean = 3.2, sd = 2.7). Ratings of mental demand and frustration were positively correlated (Spearman's $\rho = 0.442$, $p < 0.001$). In sum, the Double-or-Nothing task is a relatively brief experimental paradigm that participants find easy to comprehend, of moderate mental effort, and not unduly frustrating.