Journal of Neuroscience Methods 341 (2020) 108777

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Probing relationships between reinforcement learning and simple behavioral strategies to understand probabilistic reward learning

Eshaan S. Iyer[a,1], Megan A. Kairiss[b,1], Adrian Liu[c], A. Ross Otto[b], Rosemary C. Bagot[b,d,*]

[a] Integrated Program in Neuroscience, McGill University, 3801 Rue University, Montréal, QC H3A 2B4, Canada
[b] Department of Psychology, McGill University, 1205 Ave Dr. Penfield, Montréal, QC H3A 1B1, Canada
[c] Department of Physics, McGill University, 3600 Rue University, Montréal, QC H3A 2T8, Canada
[d] Ludmer Centre for Neuroinformatics and Mental Health, 3661 Rue University, Montréal, QC H3A 2B3, Canada

ARTICLE INFO

Keywords:
Win stay/Lose shift
Reinforcement learning
Maximum likelihood
Parameter estimation
Behavioral strategy

ABSTRACT

Background: Reinforcement learning (RL) and win stay/lose shift model accounts of decision making are both widely used to describe how individuals learn about and interact with rewarding environments. Though mutually informative, these accounts are often conceptualized as independent processes and so the potential relationships between win stay/lose shift tendencies and RL parameters have not been explored.

New method: We introduce a methodology to directly relate RL parameters to behavioral strategy. Specifically, by calculating a truncated multivariate normal distribution of RL parameters given win stay/lose shift tendencies from simulating these tendencies across the parameter space, we maximize the normal distribution for a given set of win stay/lose shift tendencies to approximate reinforcement learning parameters.

Results: We demonstrate novel relationships between win stay/lose shift tendencies and RL parameters that challenge conventional interpretations of lose shift as a metric of loss sensitivity. Further, we demonstrate in both simulated and empirical data that this method of parameter approximation yields reliable parameter recovery.

Comparison with existing method: We compare this method against the conventionally used maximum likelihood estimation method for parameter approximation in simulated noisy and empirical data. For simulated noisy data, we show that this method performs similarly to maximum likelihood estimation. For empirical data, however, this method provides a more reliable approximation of reinforcement learning parameters than maximum likelihood estimation.

Conclusions: We demonstrate the existence of relationships between win stay/lose shift tendencies and RL parameters and introduce a method that leverages these relationships to enable recovery of RL parameters exclusively from win stay/lose shift tendencies.

1. Introduction

The ability to learn from experience is crucial for survival. To successfully negotiate their environment, individuals must learn from outcomes while balancing exploration with exploitation to efficiently maximize rewards, avoid punishments, and integrate new contingencies. Learning is hypothesized to occur within a theoretical framework known as reinforcement learning (RL) in which learning accrues from the discrepancy between an expectation and an outcome, termed a prediction error (Rescorla and Wagner, 1972). Recent years have seen a surge of interest within the field of behavioral neuroscience in applying RL models to probe fundamental questions about how individuals learn (Bathellier et al., 2013; Gustafson & Daw, 2011; Kuchibhotla et al., 2019; Langdon et al., 2019; Noworyta-Sokolowska et al., 2019; Stachenfeld et al., 2017). Within the RL framework, an 'action value' is learned through prediction errors that are modulated by a learning rate, which weights the influence of prediction errors on yielding new action values. These action values are then transformed into actions using a choice rule such as the 'softmax' rule. Using RL to model decision-making behavior within an environment that yields both positive and negative outcomes can reveal differences in how individuals learn from their environment (Langdon et al., 2019; Noworyta-Sokolowska et al., 2019; St-Amand et al., 2018; Verharen et al., 2019).

* Corresponding author at: Department of Psychology, McGill University, 1205 Ave Dr. Penfield, Montréal, QC H3A 1B1, Canada.
E-mail address: rosemary.bagot@mcgill.ca (R.C. Bagot).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jneumeth.2020.108777
Received 12 December 2019; Received in revised form 22 April 2020; Accepted 11 May 2020
Available online 15 May 2020

While RL models are commonly implemented in binary choice tasks such as a probabilistic reversal learning task, they are highly flexible and can be implemented in a wide range of behavioral tasks including multi-arm bandit tasks as well as in more naturalistic foraging and spatial navigation tasks (Gustafson and Daw, 2011; Langdon et al., 2019; Noworyta-Sokolowska et al., 2019; Stachenfeld et al., 2017). At its core, RL is best suited to provide a model of how individuals interact with their environment, whether that be the rate at which they learn from feedback, internal representations of action values, or choice transition probabilities. RL models generate a variety of metrics including learning rates, choice stochasticity, action values for a given choice at a given trial, and probabilities for action selection (Daw, 2011). These parameters are estimated from a dataset by fitting a set of parameters that best explain the observed choice behavior, and can be expanded to include any number of factors the experimenter hypothesizes may be influencing behavior, such as perseveration or other choice biases. One common approach to RL model parameter estimation relies upon maximum likelihood estimation (MLE) to maximize a likelihood function to recover the most probable set of parameters given an observed choice of actions and outcomes.

Behavior within an RL framework can be understood more intuitively as following the logic of Thorndike's law of effect where the results of an action act to alter the strength of the action itself (Thorndike, 1927). RL learning rules assume that the most recent outcome exerts the most influence on the current choice, which assumes that behavior on a given trial, n, depends upon the choice and outcome on preceding trial, n − 1. This relationship has been formalized as a simple strategy known as win stay/lose shift (WSLS; (Herrnstein, 2000). This framework captures whether the previous trial was rewarded or not and if the choice on the current trial repeats or switches from the previous choice (Dalton et al., 2014). For example, if the individual is rewarded on trial n − 1 and then chooses the same option on trial n, this is termed a 'win stay', whereas if a different choice is made on trial n following reward on n − 1, this is termed a 'win shift'. If the individual is not rewarded on trial n − 1 and makes the same choice on trial n, this is termed 'lose stay', and a 'lose shift' if a different option is chosen. The proportion of win stay and lose shift responses made by an individual is interpreted as a metric of sensitivity to positive and negative outcomes, respectively. Higher win stay probabilities are interpreted as increased sensitivity to positive feedback while higher lose shift scores are interpreted as increased sensitivity to negative feedback (St Onge et al., 2011).

Both RL models and WSLS analyses attempt to describe how individuals learn from their environment, yet these approaches have largely existed in parallel literatures. A number of studies have contrasted RL and WSLS as separate strategies that individuals may differentially utilize, or as potential population markers (Ahn et al., 2014; Otto et al., 2011; Worthy and Maddox, 2012, 2014). A previous attempt to harmonize RL and WSLS into a singular model, called the WSLS-RL model, treated WSLS and RL as independent descriptors of behavior weighted differently by subject. In this model, action probabilities are first computed using a typical RL modeling with a softmax rule and are then 'mixed' with WSLS probabilities to generate new action probabilities (Worthy and Maddox, 2014). However, if RL is an accurate description of learning and behavior, WSLS and RL analyses are likely not independent descriptors of behavior. Though WSLS uses a probabilistic description of strategy-based tendencies while RL uses the concept of learning rates to propagate prediction errors, both WSLS and RL rely heavily upon recent outcomes to explain future behavior. This suggests that these models describe some sort of common underlying proclivity, leading us to predict a correspondence between WSLS tendencies and RL parameters.

Here, we demonstrate, with both model simulation exercises and empirical data, that RL parameters and WSLS tendencies can, under certain circumstances, serve as mutually informative descriptors of behavior. To do this, we examine the relationship between WSLS tendencies and RL parameters in a probabilistic reinforcement task in both simulated data and rodent behavioral data. We demonstrate how this relationship can be used to calculate a normal distribution function that can be maximized to estimate RL parameters, under certain circumstances, on the basis of WSLS tendencies. Because WSLS is a behavioral tendency integrated over many trials, it should not change dramatically with small fluctuations or perturbations in behavior. We suggest that this method of RL parameter recovery is robust in data sets with noise as well as small numbers of trials, features common to behavioral neuroscience research. Furthermore, our novel WSLS parameter estimation approach renders RL modeling readily accessible to behavioral neuroscience researchers who may not have specific computational modeling expertise, allowing for simplified recovery of RL parameters from input of WSLS probabilities.

2. Methods

2.1. Model simulations and estimation

2.1.1. Task specification

2.1.1.1. Probabilistic binary choice task. In this probabilistic binary choice task, an agent was required to choose between one of two options with one option delivering a reward with 80% probability and the other option rewarded at 20% probability. Agents were required to make either 100 or 1000 total choices depending on the specified model.

2.1.1.2. Probabilistic reversal learning task. This probabilistic reversal learning task followed the same basic conditions described above in the probabilistic binary choice task with the added complexity that following five consecutive choices of the higher reward probability option, the contingencies reversed such that the option that had delivered a reward at 80% probability now delivered reward with 20% probability and the option that had delivered a reward at 20% probability now delivered a reward with 80% probability (Dalton et al., 2014; St Onge et al., 2011; Verharen et al., 2019). The number of reversals was not constrained other than by the total number of choices required (100 or 1000 total choices depending on the specified model).

2.1.2. Reinforcement learning model

We simulated the behavior of a reinforcement learning agent that utilizes the model-free reinforcement learning approach in a probabilistic binary choice task and a probabilistic reversal learning task (Rescorla and Wagner, 1972; Sutton and Barto, 2011; Watkins, 1989). In this model, an agent selects a certain option on trial t that results in the delivery of a reward or not. For each trial, the expected reward value, or Q value, for an action a_i is compared with the actual reward, yielding a prediction error δ_t,

$$\delta_t = r_t - Q(a_i, t)$$

where $Q(a_i, t)$ represents the expected reward value for an action a_i and r_t represents whether or not a reward was delivered $r_t \in [0,1]$. This prediction error is modified by the learning rate parameter (α) and used to update the Q values for action a_i for the following trial where α is bounded between 0 and 1.

$$Q(a_i, t+1) = Q(a_i, t) + \alpha \cdot \delta_t$$

To calculate the probability of choosing a certain action, the Q values are inputted into a softmax decision rule, multiplied by an 'inverse temperature' parameter (β; bounded between 0 and 10):

$$P(a_i) = \frac{\exp(\beta \cdot Q(a_i, t))}{\sum_{j=1}^{2} \exp(\beta \cdot Q(a_j, t))} \quad (3)$$

As the value of β tends towards infinity, the highest Q value option is more likely to be chosen while as β tends towards 0, both options

become equally probable.

2.1.3. Win stay lose shift (WSLS)

We calculated win stay/lose shift tendencies from the reinforcement learning agent's behavior in the probabilistic binary choice and probabilistic reversal learning tasks. To do this we defined a win stay (WS) as trials in which the previous trial (n − 1) was rewarded and the choice on the current trial n is the same with the choice on trial n − 1. We calculated WS probabilities as the proportion of trials with WS behavior given a previously rewarded trial. We defined a lose shift (LS) as a trial in which the previous trial (n − 1) was unrewarded and the choice on the current trial n differs from the choice on trial n − 1. We calculated LS probabilities as a proportion of trials with LS behavior given that the previous trial was unrewarded.

Following the assumptions made by RL, recent outcomes have a greater influence on a current choice. To account for the influence of recent but not immediate outcomes for a given choice, we also examined the influence of 2-back trials (n − 2) on choice on current trial n by looking at 2-back WSLS tendencies in conjunction with 1-back trials (n − 1). To do this we defined win stay-2 (WS2) as being trials in which trial n − 2 was rewarded and the choice on the current trial n is concordant with the choice on trial n − 2. We calculated WS2 probabilities as a proportion of trials with WS2 behavior given that the trial n − 2 was rewarded. We defined lose shift-2 (LS2) as being trials in which trial n − 2 was unrewarded and the choice on the current trial n differs from the choice on trial n − 2. We calculated LS2 probabilities as a proportion of trials with LS2 behavior given that the trial n − 2 was unrewarded.

2.1.4. WSLS estimation (WSLSE)

To establish the relationship between WSLS tendencies and RL, we simulated behavior in a learning task using a RL model. We discretized the parameter space to include α values between 0.01 and 1, incremented linearly in steps of 0.01, and β between 0 and 10 incremented linearly in steps of 1. For each possible combination of α and β, we simulated trial-by-trial behavior for 100 and for 1000 trials using the RL algorithm detailed above with the starting Q values initialized to 0.5.

We used this simulated choice behavior to approximate a multivariate probability distribution function to calculate the set of RL parameters that correspond to a given set of WSLS behavior. To do this we first calculated WS, LS, WS2, and LS2 probabilities at each combination of RL parameters. We repeated this process 1000 times for each parameter combination to account for the probabilistic nature of the simulation process. We then calculated the variance of WS, LS, WS2, and LS2 using Eq. (4),

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{4}$$

and the covariance between WS, LS, WS2, and LS2 using Eq. (5)

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{5}$$

at each combination of RL parameters. Using the variances and covariance of WSLS tendencies we then created a series of covariance matrices: $\Sigma_{WS,\,LS\,(100)}$, for WS and LS behavior for 100 trials, $\Sigma_{WS,\,LS\,(1000)}$, for WS and LS behavior for 1000 trials, $\Sigma_{WS,\,LS,\,WS2,\,LS2\,(100)}$, for WS, LS, WS2, and LS2 behavior for 100 trials, and $\Sigma_{WS,\,LS,\,WS2,\,LS2\,(1000)}$, for WS, LS, WS2, and LS2 behavior for 1000 trials. We used either 100 or 1000 trials, as these are representative of trial numbers typically seen experimentally. These four unique covariance matrices allowed us to independently examine how increasing trial number and adding a 2-back component to the model impacts accuracy.

Because WSLS tendencies lie between 0 and 1, we used a truncated multivariate normal distribution to model the observed multivariate WSLS distribution at each given combination of RL parameters. To accomplish this, we utilized the tmvnorm package in R (Wilhelm and Manjunath, 2010), which gives the generalized case of the truncated multivariate Gaussian as

$$f(x, \mu, a, b) = \frac{\exp[-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)]}{\int_a^b \exp[-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)]dx} \tag{6}$$

where x represents a vector of observational variables, which in our case is either [WS, LS] or [WS, LS, WS2, LS2], μ represents the vector of means, a represents the vector of upper truncation points, and b represents the vector of lower truncation points. In correspondence with the upper and lower bounds of WS and LS tendencies, we truncated WS, LS, WS2, and LS2 at 0 and 1. Using Eq. (6), we approximated a normal probability distribution function for the RL parameter space, given a set of WSLS values. By maximizing this probability distribution function, we are able to identify a set of RL parameters that best describes the observed WSLS behavior.

This code is accessible through https://github.com/BagotLab/WSLSE.

2.1.5. Maximum likelihood estimation (MLE)

To fit RL parameters to data, we used the Nelder-Mead (1965) optimization method to identify parameters that maximized the likelihood of an agent's choice given its history of choices and outcomes (cf. Daw, 2011).

2.2. Empirical data analysis

2.2.1. Animals

Seven-week-old C57/Bl6J mice (12 male, 12 female) were obtained from Jackson Laboratories and housed in same-sex groups of four on a regular 12 h light-dark cycle at 22 − 25 °C with ad libitum access to food and water for one week prior to behavioral training. At the start of training, food was removed and mice were food restricted to 85% of their original body weight. Animals were weighed daily throughout the course of the experiment and food adjusted as required to maintain bodyweight. All experiments were conducted in accordance with guidelines of McGill's Animal Care Committee.

2.2.2. Behavioral testing

Operant conditioning was conducted in sound attenuating conditioning chambers (Med Associates) with two retractable levers either side of a food port from which 20 mg chocolate pellets (Bio-Serv Inc.) were delivered. Animals were initially trained on a fixed ratio-1 (FR1) schedule in which each press on the designated active lever resulted in the delivery of a chocolate pellet followed by a five second time-out period during which further lever presses had no effect. Presses on the other lever, designated the inactive lever, at any point had no effect. Following three days of FR1, animals progressed to the Probabilistic Binary Choice Task in which one lever was rewarded with a probability of 80% and the other lever was rewarded with a probability of 20%. Following each lever press, both levers retracted for five seconds before extending again to indicate trial onset. Each training session lasted 30 min.

2.2.3. WSLS estimation (WSLSE)

For approximation of RL parameters from empirical animal behavioral data, WSLSE was performed on the first 100 trials following the same procedure as used for data simulations, outlined above. Briefly, for each animal, we approximated and then maximized the probability distribution function for a set of RL parameters using the WSLSE method. Win stay/lose shift probabilities (WS, LS, WS2, LS2) were calculated based on the action history of each animal. The means and covariance matrices used were generated from the simulated choice behavior described above.

2.2.4. Maximum likelihood estimation (MLE)

To fit RL parameters from empirical animal behavioral, we used the Nelder-Mead optimization on the first 100 trials as described above (Nelder and Mead, 1965).

2.2.5. Statistics

Inferential statistical analyses were performed using Prism 7 (GraphPad Software Inc.). To compare parameter estimates in empirical data, we compared group means by *t*-test and variances by an F-test of equality of variances. Grubbs' test for outliers was used in the empirical data analysis to identify and exclude outliers.

3. Results

3.1. WSLS tendencies and RL parameters

To probe the relationship between WSLS tendencies and RL parameters, we simulated the behavior of an RL agent that uses a single learning rate in a simple probabilistic binary choice task where one option is rewarded with 80% probability and the other option is rewarded with 20% probability. For each combination of learning rate and inverse temperature parameter, we simulated this agent's behavior 1000 times to calculate mean win stay and lose shift probabilities. Fig. 1A depicts win stay probabilities as a function of the learning rate and inverse temperature parameters for a range of plausible parameter values. For a given inverse temperature parameter value, win stay probabilities increase as the learning rate increases before reaching a plateau. Likewise, for a given learning rate, increasing the inverse temperature parameter, increases win stay probabilities, though in a more dramatic way. This relationship tracks with our understanding of RL. Individuals with high learning rates learn more from prediction errors and thus, following a rewarding outcome, are more likely to stick with the previously rewarded choice.

The observed relationship between lose shift tendencies and RL parameters (Fig. 1C) is more complex. At very low inverse temperatures, regardless of learning rate, lose shift probabilities do not deviate noticeably from 0.5. At higher inverse temperature parameters—meaning choices are more sensitive to learned values—the relationship between lose shift probabilities and learning rate assumes an inverted-U shape, with lose shift probabilities first decreasing then subsequently increasing with further increases in learning rate. The minimum of this inverted-U decreases as the inverse temperature increases. Lose shift peaks at both low and high learning rates with the lowest lose shift probabilities observed at intermediate learning rates. The U-shaped relationship is likely due to the fact that at very low learning rates, agents learn little from either reward or loss causing them to behave more randomly. As an agent begins to learn more from reward and loss, we see their behavior quickly drop off to low lose shift probabilities at slightly higher, but still relatively low learning rates, and then increase as the learning rates increase. For both win stay and lose shift, as inverse temperature decreases, the probability of behavior approaches chance. A similar effect can be seen for win stay and lose shift probabilities as learning rate decreases, but this effect appears to be more strongly modulated by inverse temperature. We also probed the relationship between 2-back trial behavior and RL parameters to examine the influence of recent, but not immediate trials. As expected, we find the relationship between 2-back win stay and RL parameters is similar but not identical to that of the 1-back win-stay (Fig. 1B) and the relationship between 2-back lose shift and RL parameters is similar to that of 1-back lose shift (Fig. 1D).

The relationship between win stay and lose shift tendencies also appears to vary with RL parameters across both 1- (Fig. 1E) and 2-back trials (Fig. 1F). In both 1- and 2-back trials, this relationship takes a similar inverted-U shape to that between lose shift tendencies and RL parameters. At low inverse temperatures, win stay and lose shift tendencies are very similar with this ratio decreasing as inverse temperature increases and decreasing, then increasing, as learning rate increases.

Having established the relationships between RL parameters and win stay and lose shift tendencies, we then reasoned that, if indeed these relationships exist, it should be possible to predict RL parameters given a set of win stay and lose shift probabilities by approximating a multivariate probability distribution of RL parameters for a given set of win stay and lose shift probabilities. To accomplish this, we calculated and then maximized a truncated multivariate normal distribution of RL parameters given specific win stay and lose shift probabilities.

3.2. WSLSE for RL parameter recovery

To assess the accuracy of parameter recovery by win stay lose shift estimation (WSLSE), we correlated recovered RL parameters to simulated ground truths established by randomly sampling a uniform distribution of learning rates between 0 and 1 and inverse temperatures between 0 and 10 (Ballard and McClure, 2019; Virtanen et al., 2020; Wilson and Collins, 2019). Each randomly sampled RL parameter was then used to simulate 100 choices in a fixed binary choice task from which win stay and lose shift probabilities were calculated. We selected a set size of 100 choices as it is representative of trial set sizes commonly used to estimate RL parameters in the literature (St-Amand et al., 2018; Wimmer et al., 2014; Wunderlich et al., 2009). We then used the win stay and lose shift probabilities to calculate then maximize a truncated multivariate normal distribution to estimate the RL parameters. Repeating this process 1000 times, we correlated the means of the truncated normal distributions to the randomly sampled RL parameters. The correlation between ground-truth parameter values and parameters estimated by WSLSE for 100 trials was $r = 0.7188$ for learning rate (Fig. 2A) and $r = 0.8004$ for inverse temperature (Fig. 2C). This increased to $r = 0.9204$ for learning rate (Fig. 2E) and $r = 0.9628$ for inverse temperature (Fig. 2G) for 1000 trials. In comparison, the conventionally used maximum likelihood estimation (MLE) approach yielded a correlation to ground truth of $r = 0.7734$ for learning rate (Fig. 2B) and $r = 0.8383$ for inverse temperature (Fig. 2D) for 100 trials. For 1000 trials, this increased to $r = 0.9379$ for learning rate (Fig. 2F) and $r = 0.9707$ for inverse temperature (Fig. 2H). The average negative log likelihood of parameters recovered using WSLSE are similar to those estimated using MLE (Table 1). The negative log likelihood values and correlations suggest that this WSLSE method can produce reasonably accurate parameter recovery comparable to the conventionally used MLE method, and also confirms the relationship we identified between WSLS tendencies and RL parameters.

Visually comparing correlations between WSLSE and MLE, suggests that MLE reacts differently than WSLSE when approximating RL parameters. When MLE struggles to describe the data, parameters are pushed towards the boundaries (Daw, 2011). In contrast, under similar conditions, WSLSE confines parameters within a restricted range constrained by the win stay and lose shift probabilities, suggesting that, while WSLSE may be somewhat less precise, in real data it may be more accurate than MLE and less prone to dramatic errors. Visual examination of these correlations also illustrates that at low trial numbers WSLSE has a tendency to bias away from low learning rates. This suggests that in conditions where experimenters have reason to expect very low learning rates, MLE may be preferable to WSLSE.

3.3. -Back WSLSE for RL parameter recovery

A key assumption made by RL models, by virtue of the learning rules employed (Eqs. (1) and (2)), is that more recent outcomes will exert greater influence on a current choice. However, to account for the influence of recent trials on the current trial beyond 1 trial back, we incorporated win stay and lose shift probabilities from 2-back trials alongside the win stay and lose shift probabilities from 1-back trials. These two additional observational variables further restrict the range

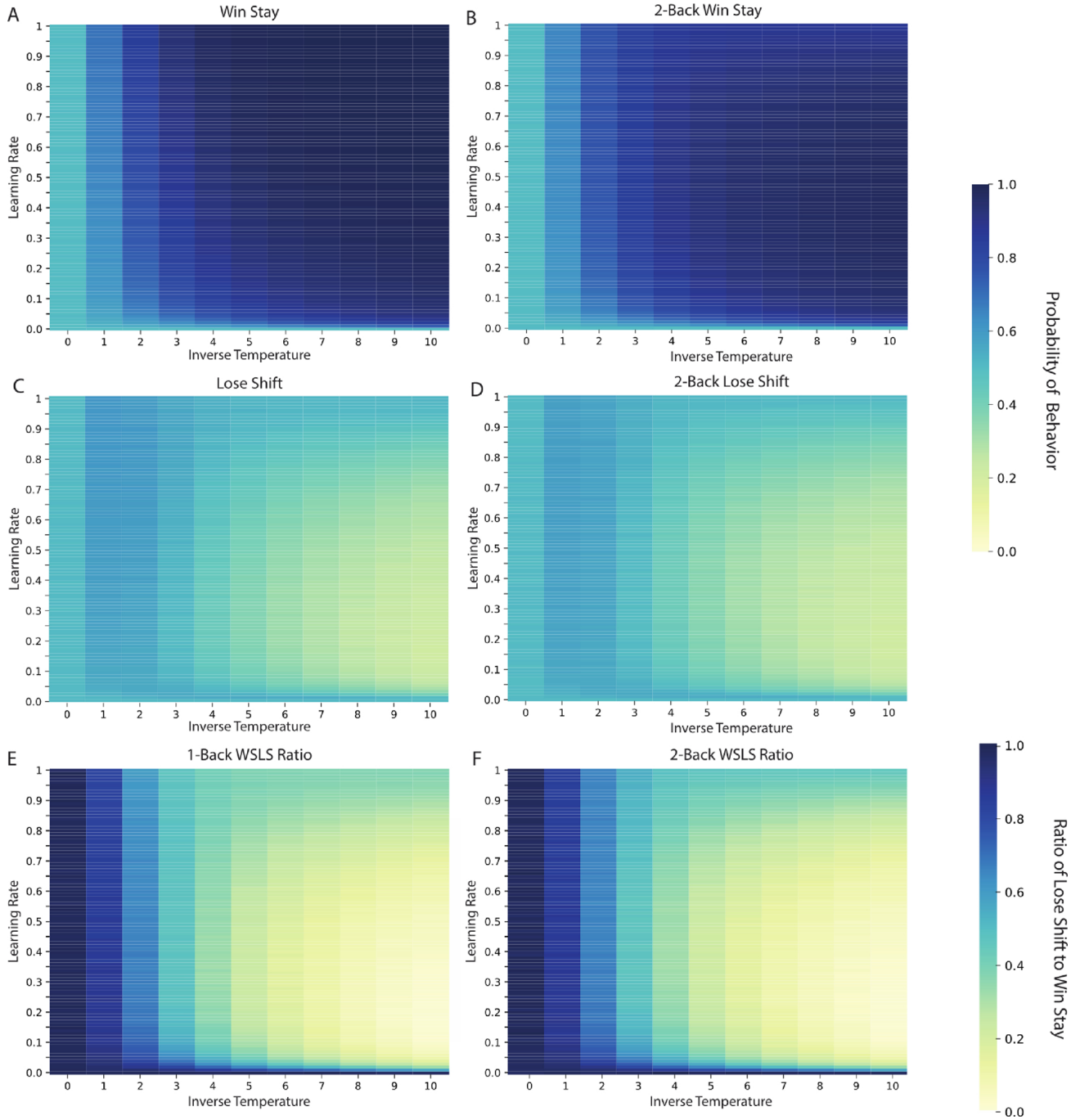

Fig. 1. (a) Heatmap visualization of simulated mean win stay behavior as a function of RL parameters in a probabilistic binary choice task. Holding the inverse temperature constant and increasing the learning rate, WS probabilities increase before reaching a plateau. A similar relationship is observed increasing the inverse temperature across a constant learning rate. (b) Heatmap visualization of simulated mean 2-back win stay behavior. 2-back win stay behavior is similar to 1-back win stay behavior, however, increasing the inverse temperature for high learning rates does not increase WS probabilities as much. (c) Heatmap visualization of simulated mean lose shift behavior as a function of RL parameters. Although complex, in general, as inverse temperature increases, LS tends to decrease, and tends towards a minimum at intermediate learning rates and a maximum at low and high learning rates. (d) Heatmap visualization of simulated mean 2-back lose shift behavior. 2-back lose shift behavior is similar to 2-back lose shift behavior with higher probabilities closer to 0.5. (e) Heatmap visualization of the ratio of simulated mean lose shift behavior to win stay behavior as a function of RL parameters in a probabilistic binary choice task. A ratio closer to 1 indicates equal probabilities of win stay and lose shift behavior while a ratio closer to 0 indicates a higher probability of win stay behavior compared to lose shift behavior. (f) Heatmap visualization of the ratio of simulated mean 2-back lose shift behavior to 2-back win stay behavior as a function of RL parameters in a probabilistic binary choice task. A ratio closer to 1 indicates equal probabilities of win stay and lose shift behavior while a ratio closer to 0 indicates a higher probability of win stay behavior compared to lose shift behavior.

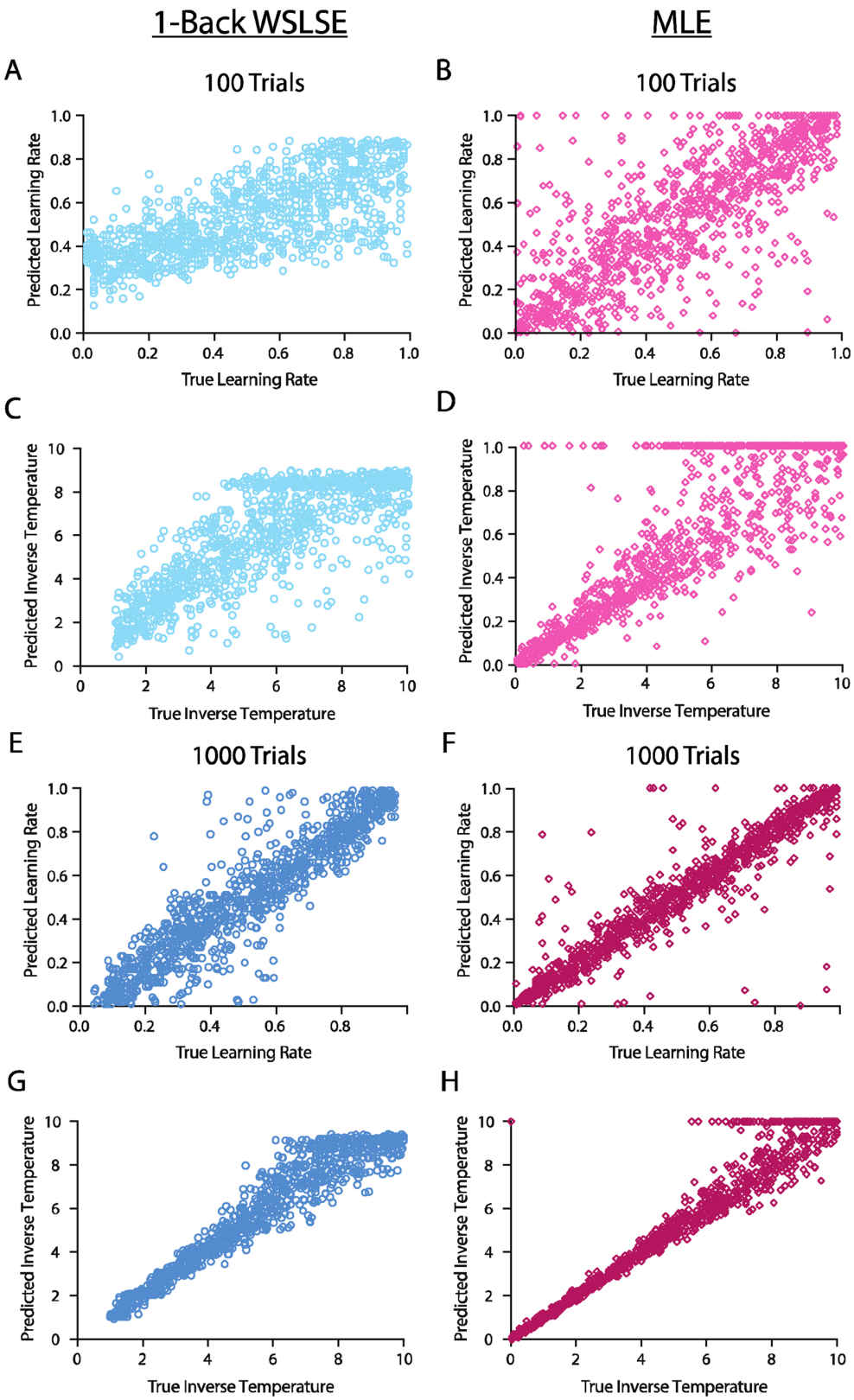

Fig. 2. Correlation between simulated ground truth and parameters recovered in a Probabilistic Binary Choice Task for (a) learning rate ($r = 0.7188$, $p < 0.0001$) using 1-back WSLSE, (b) learning rate ($r = 0.7734$, $p < 0.0001$) using MLE, (c) inverse temperature ($r = 0.8004$, $p < 0.0001$) using 1-back WSLSE, (d) inverse temperature ($r = 0.8383$, $p < 0.0001$) using MLE for 100 trials for 1000 simulated subjects. Correlation between simulated ground truth and parameters recovered in a Probabilistic Binary Choice Task for (e) learning rate ($r = 0.9204$, $p < 0.0001$) using 1-back WSLSE, (f) learning rate ($r = 0.9379$, $p < 0.0001$) using MLE, (g) inverse temperature ($r = 0.9628$, $p < 0.0001$) using 1-back WSLSE, (h) inverse temperature ($r = 0.9707$, $p < 0.0001$) using MLE for 1000 trials for 1000 simulated subjects.

Table 1
Average negative log-likelihood values for parameters recovered using 1-Back Win stay Lose shift Estimation (WSLSE), 2-Back WSLSE, and Maximum Likelihood Estimation (MLE) for 100 and 1000 trials in a Probabilistic Binary Choice Task.

	1-Back WSLSE	2-Back WSLSE	MLE
100 trials	30.085	29.490	27.895
1000 trials	250.654	250.241	247.448

of possible RL parameters and so should serve to increase parameter recovery. We used a similar procedure as 1-back WSLSE, using win stay, lose shift, 2-back win stay, and 2-back lose shift probabilities to calculate and then maximized a truncated multivariate normal distribution to estimate RL parameters. Correlating simulated ground truths to recovered parameters for 2-back WSLSE for 100 trials yielded a correlation of r = 0.7752 for learning rate (Fig. 3A) and r = 0.8055 for inverse temperature (Fig. 3C). Using MLE for 100 trials, this resulted in a correlation of r = 0.8009 for learning rate (Fig. 3B) and r = 0.8383 for inverse temperature (Fig. 3D). For 1000 trials this yielded a correlation of r = 0.9533 for learning rate (Fig. 3E) and r = 0.9629 for inverse temperature (Fig. 3G). Using MLE for 100 trials, this resulted in a correlation of r = 0.9457 for learning rate (Fig. 3F) and r = 0.9756 for inverse temperature (Fig. 3H).

Comparing the negative log likelihoods for the 1-back WSLSE and 2-back WSLSE reveals that at 100 trials, 2-back WSLSE has a lower negative log likelihood, indicating higher likelihood (Table 1). The negative log likelihood is also marginally lower for 2-back WSLSE at 1000 trials, mirroring the marginal improvement in accuracy of 2-back WSLSE over at 1000 trials. Thus while 2-back WSLSE for RL parameter recovery modestly improves accuracy for 100 trials and increases the likelihood of recovered parameters, it offers no clear advantage over 1-back WSLSE for 1000 trials.

3.4. WSLSE is applicable in a range of tasks

To illustrate the adaptability and broader applicability of WSLSE, we assessed parameter recoverability in a reversal learning task. We simulated the behavior of an agent that uses a single learning rate in a probabilistic reversal learning task where one option is rewarded with 80% probability and the other option 20% and the contingencies reverse after choosing the higher reward probability option on five consecutive trials (St Onge et al., 2011; Verharen et al., 2019). To confirm the existence of relationships between behavior in the reversal learning task and RL parameters, we plotted heatmaps of simulated mean win stay and lose shift behavior. Though similar to behavior in the probabilistic binary choice task, we find lose shift behavior is generally more probable in this reversal learning task (Fig. 4). We then calculated a covariance matrix for win stay and lose shift behavior and used this to calculate then maximize a truncated multivariate normal distribution of RL parameters given specific win stay, lose shift, 2-back win stay, and 2-back lose shift probabilities. Correlating ground-truth parameter values to recovered parameters within a probabilistic reversal learning task yielded a correlation of r = 0.7342 for learning rate (Fig. 5A) and r = 0.8158 for inverse temperature (Fig. 5C) for 100 trials. In simulations with 1000 trials, this correlation increased to r = 0.9535 for learning rate (Fig. 5E) and r = 0.9351 for inverse temperature (Fig. 5G). Using MLE, this yielded a correlation to ground truth of r = 0.7830 for learning rate (Fig. 5B) and r = 0.8426 for inverse temperature (Fig. 5D) for 100 trials and a correlation of r = 0.9457 for learning rate (Fig. 5H) and r = 0.9756 for inverse temperature (Fig. 5F) for 1000 trials. As with the probabilistic binary choice task, the average negative log likelihood of parameters recovered using WSLSE are comparable to those estimated using MLE (Table 2). This demonstrates that WSLSE can recover parameters for a range of tasks, requiring only the simulation of agent behavior within a specified task and the calculation of a corresponding covariance matrix for WSLS behavior.

3.5. WSLSE for noisy RL parameter recovery

Classical simulations likely overestimate parameter recoverability as they assume choice behavior is determined solely by RL processes, and fail to consider the inherently noisy and idiosyncratic nature of observed behavior. To account for this, we repeated the above correlation exercise, modeling one of three sources of systematic noise in the 2-back model for 100 trials. We used the 2-back model to assess learning rate recovery under noise because of the improved accuracy it offers at lower trial numbers as demonstrated above. We compared the use of WSLSE to the conventionally used maximum likelihood estimation (MLE) approach to estimate RL parameters from noisy data. The amount of noise was scaled from zero to ten percent to determine how WSLSE compares to MLE for increasingly noisy data. To model random, non-systematic noise, trials were 'flipped'. Changing a small percent of trials should not significantly shift RL parameters and so a reliable method of estimating RL parameters should be robust to such noise at low levels. To model perseverance-derived noise (i.e., unexplained streaks of consecutive choices of one action), a sequence of singular responses on consecutive trials was inserted. This type of noise is often observed in behavioral data and may represent a distinct perseverative process occurring alongside learning. To model noise derived from an alternating strategy, a sequence of alternating choices was inserted randomly into each simulated data set. Similar to perseverative noise, this type of noise is observed in behavioral data and is suggestive of a distinct strategy-based approach that may occur alongside learning processes. Because the goal of RL approaches is to describe the learning processes that underlie behavior, a good RL model should be robust to such noise.

We found that the method was reasonably robust to random noise; flipping the results of an increasing percent of 100 trials resulted in relatively stable parameter recovery for both MLE and WSLSE across both learning rate and inverse temperature with slightly higher accuracy using MLE to recover learning rate (Fig. 6A and B). WSLSE was largely robust to perseverance-derived noise, performing similarly to MLE for learning rate recovery (Fig. 6D). Interestingly, WSLSE was more robust than MLE for inverse temperature parameter recovery, remaining generally stable as perseverance-derived noise increased (Fig. 6E). Both MLE and WSLSE were less robust for learning rate parameter recovery in data with noise from an alternating strategy (Fig. 6G). WSLSE for inverse temperature parameter recovery, however, was more robust to noise from an alternating strategy than MLE (Fig. 6H). Structured noise arising from perseverance or alternating strategies indicate that an individual is no longer sensitive to outcomes, with choice behavior guided by some alternate mechanism. This lack of sensitivity to outcome represents a deviation from the assumptions of reinforcement learning and thus understandably decreases the accuracy of parameter recovery. MLE and WSLSE both yielded similar negative log likelihood values across the various types of noise with MLE expectedly yielding slightly higher negative log likelihood values than WSLSE (Fig. 6C, F and I).

3.6. WSLSE for analysis of empirical animal choice behavior

The utility of a learning model such as RL rests upon its ability to describe phenomena in empirical data. To assess this, we replicated the simulated probabilistic binary choice task, in a cohort of male and female mice to probe potential differences in how WSLSE and the conventionally implemented MLE describe empirical data. We first calculated 1 and 2-back win stay and lose shift behavior (Fig. 7A and B) and used these values to implement WSLSE. We did not detect any average differences between WSLSE and MLE of either learning rate or inverse temperature parameters (Fig. 7C, D and E). This demonstrates that at a

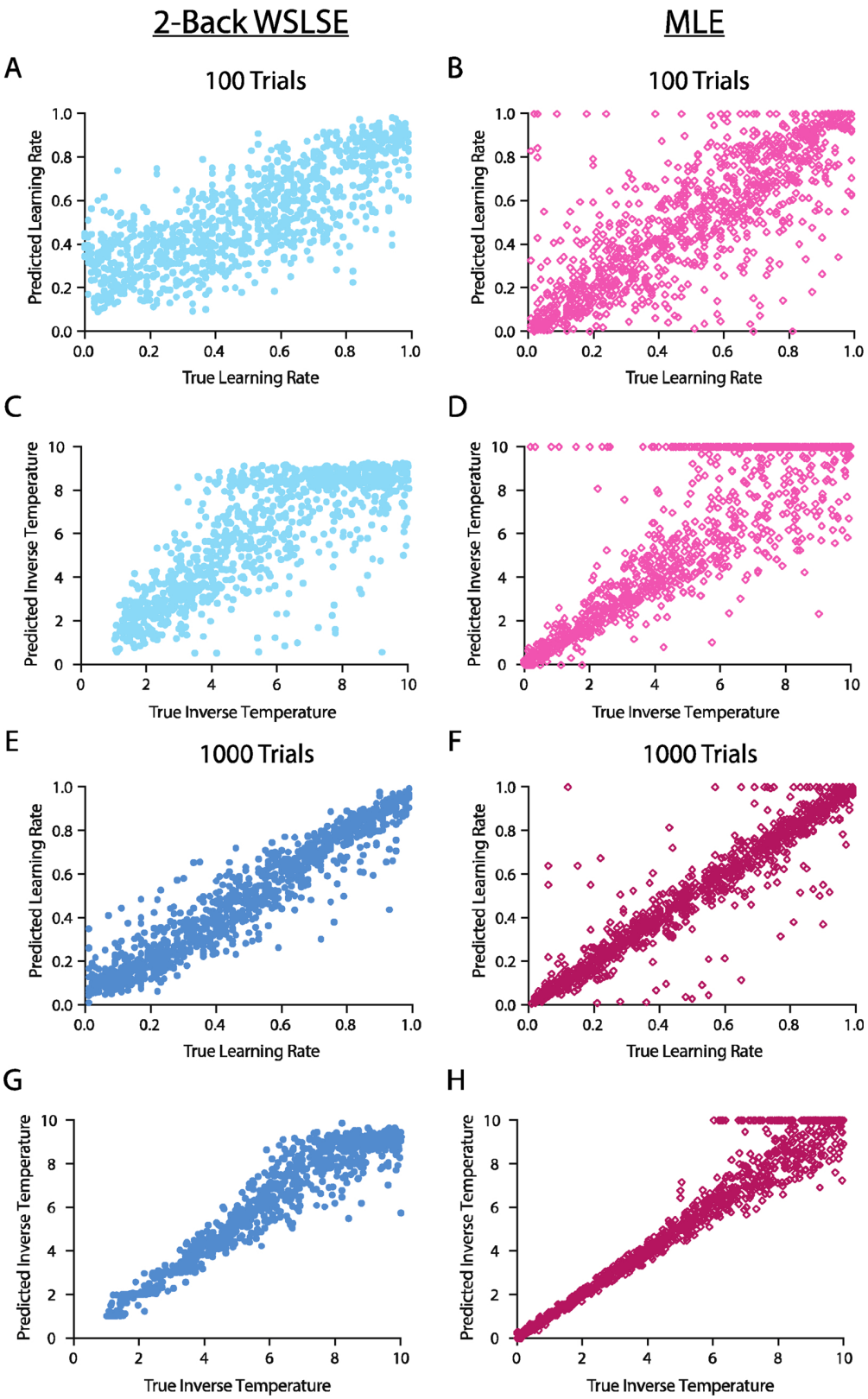

Fig. 3. Correlation between simulated ground truth and parameters recovered in a Probabilistic Binary Choice Task for (a) learning rate ($r = 0.7752$, $p < 0.0001$) using 2-back WSLSE, (b) learning rate ($r = 0.8009$, $p < 0.0001$) using MLE, (c) inverse temperature ($r = 0.8055$, $p < 0.0001$) using 2-back WSLSE, (d) inverse temperature ($r = 0.8383$, $p < 0.0001$) using MLE for 100 trials for 1000 simulated subjects. Correlation between simulated ground truth and parameters recovered in a Probabilistic Binary Choice Task for (e) learning rate ($r = 0.9533$, $p < 0.0001$) using 2-back WSLSE, (f) learning rate ($r = 0.9457$, $p < 0.0001$) using MLE, (g) inverse temperature ($r = 0.9629$, $p < 0.0001$) using 2-back WSLSE, (h) inverse temperature ($r = 0.9756$, $p < 0.0001$) using MLE for 1000 trials for 1000 simulated subjects.

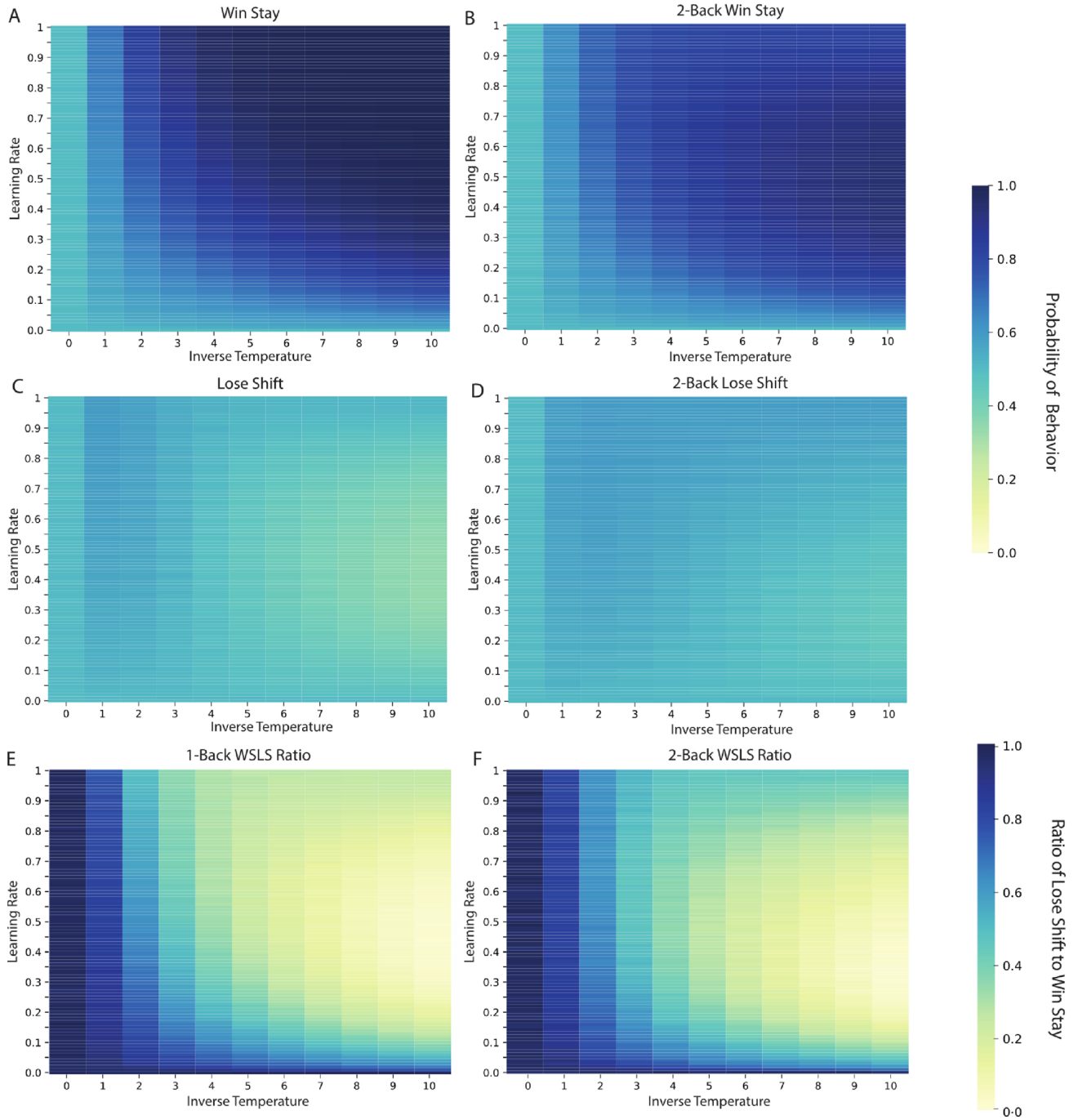

Fig. 4. (a) Heatmap visualization of simulated mean win stay behavior as a function of RL parameters in a probabilistic reversal learning task. (b) Heatmap visualization of simulated mean 2-back win stay behavior. 2-back win stay behavior is similar to 1-back behavior, though increasing the inverse temperature for high learning rates does not increase WS probabilities as much. (c) Heatmap visualization of simulated mean lose shift behavior as a function of RL parameters in a probabilistic reversal learning task. (d) Heatmap visualization of simulated mean 2-back lose shift behavior. 2-back lose shift behavior is similar to 1-back behavior with higher probabilities closer to 0.5. (e) Heatmap visualization of the ratio of simulated mean lose shift behavior to win stay behavior as a function of RL parameters in a probabilistic reversal learning task. A ratio closer to 1 indicates equal probabilities of win stay and lose shift behavior while a ratio closer to 0 indicates a higher probability of win stay behavior compared to lose shift behavior. (f) Heatmap visualization of the ratio of simulated mean 2-back lose shift behavior to 2-back win stay behavior as a function of RL parameters in a probabilistic reversal learning task. A ratio closer to 1 indicates equal probabilities of win stay and lose shift behavior while a ratio closer to 0 indicates a higher probability of win stay behavior compared to lose shift behavior.

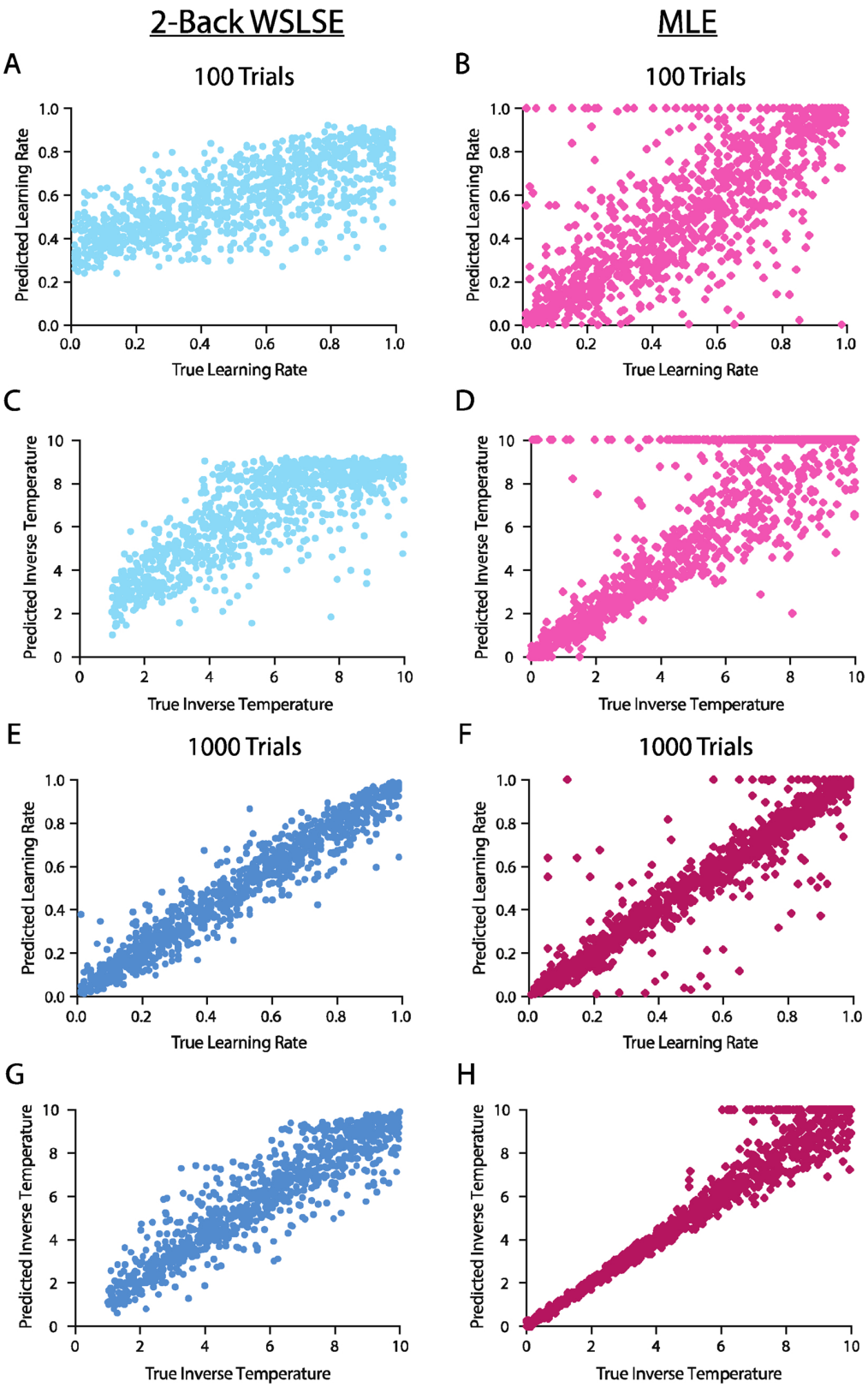

Fig. 5. Correlation between simulated ground truth and parameters recovered in a probabilistic reversal learning task for (a) learning rate ($r = 0.7342$, $p < 0.0001$) using 2-back WSLSE, (b) learning rate ($r = 0.7830$, $p < 0.0001$) using MLE, (c) inverse temperature ($r = 0.8158$, $p < 0.0001$) using 2-back WSLSE, (d) inverse temperature ($r = 0.8426$, $p < 0.0001$) using MLE for 100 trials for 1000 simulated subjects. Correlation between simulated ground truth and parameters recovered in a probabilistic reversal learning task for (e) learning rate ($r = 0.9695$, $p < 0.0001$) using 2-back WSLSE, (f) learning rate ($r = 0.9457$, $p < 0.0001$) using MLE, (g) inverse temperature ($r = 0.9351$, $p < 0.0001$) using 2-back WSLSE, (h) inverse temperature ($r = 0.9756$, $p < 0.0001$) using MLE for 1000 trials for 1000 simulated subjects.

Table 2
Average negative log-likelihood values for parameters recovered using 2-Back WSLSE, and Maximum Likelihood Estimation for 100 and 1000 trials in a probabilistic reversal learning task.

	2-Back WSLSE	MLE
100 trials	42.422	42.489
1000 trials	473.768	437.694

population level, our method of WSLSE estimation performs just as well as MLE in empirical data.

However, assessment of equality of variances indicated a significant difference between WSLE and MLE for both learning rate ($S^2_{WSLSE} = 0.00541$, $S^2_{MLE} = 0.013142$ $F_{21,21} = 2.427$, $p = 0.0481$) and inverse temperature ($S^2_{WSLSE} = 0.66073$, $S^2_{MLE} = 11.50416$, $F_{21,21} = 17.41$, $p = 0.0001$). This indicates that while both approaches arrive at similar group-level estimates, MLE approximates a more variable set of RL parameters. This is particularly apparent in the exploitation parameter estimates. Because the RL parameters are being estimated from a single population, we would not expect to see the spread of parameters observed in the MLE estimates. MLE also tends to push the exploitation parameter, β, estimates towards the boundary value of 10. This mirrors the observation in simulated data, that, when MLE struggles to describe data, it pushes parameters to the boundary (Daw, 2011). Here in the observed choice data, this is likely a consequence of the small number of trials (100). The more constrained set of parameters generated by WSLSE suggests this parameter estimation method generates more accurate parameter values than the typical MLE approach under the suboptimal conditions of lower trial numbers or increased noise that commonly occur in empirical data.

4. Discussion

Here we demonstrated the existence of a relationship between WSLS tendencies and RL and then leveraged this to create a novel method of

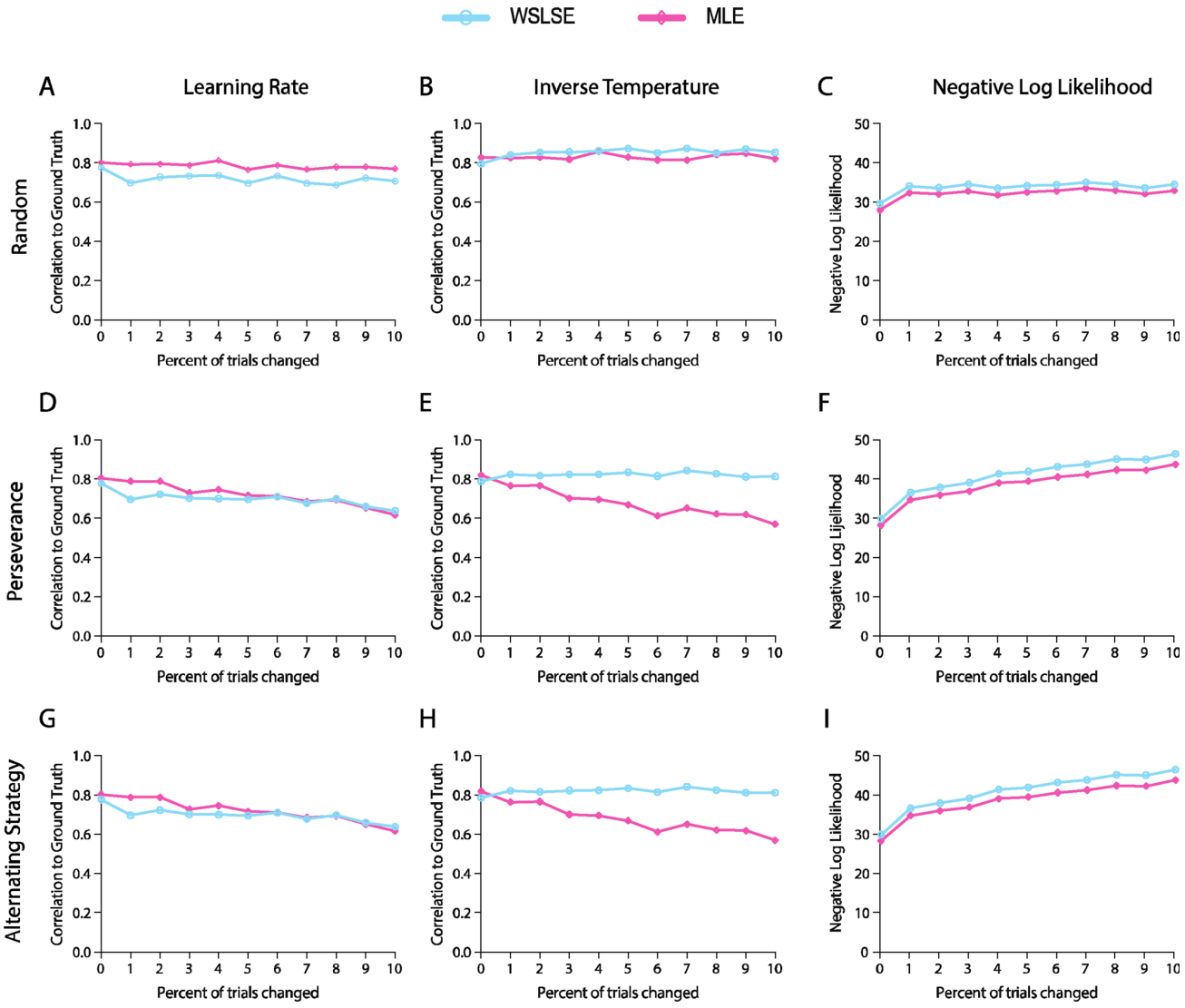

Fig. 6. Correlation between simulated ground truths and recovered parameters in a probabilistic binary choice task for 100 trials for 1000 simulated subjects using WSLSE and MLE across simulated noise varying between zero and ten percent (a) with modeled random noise for learning rate and (b) inverse temperature; (d) with modeled perseverance-derived noise for learning rate and (e) inverse temperature; (g) with modeled alternating strategy-derived noise for learning rate and (h) inverse temperature. Average negative log likelihood values for recovered parameters in a probabilistic binary choice task for 100 trials for 1000 simulated subjects recovered using WSLSE and MLE across simulated noise varying between zero and ten percent (c) with modeled random noise, (f) with modeled perseverance-derived noise, and (i) with modeled alternating strategy-derived noise.

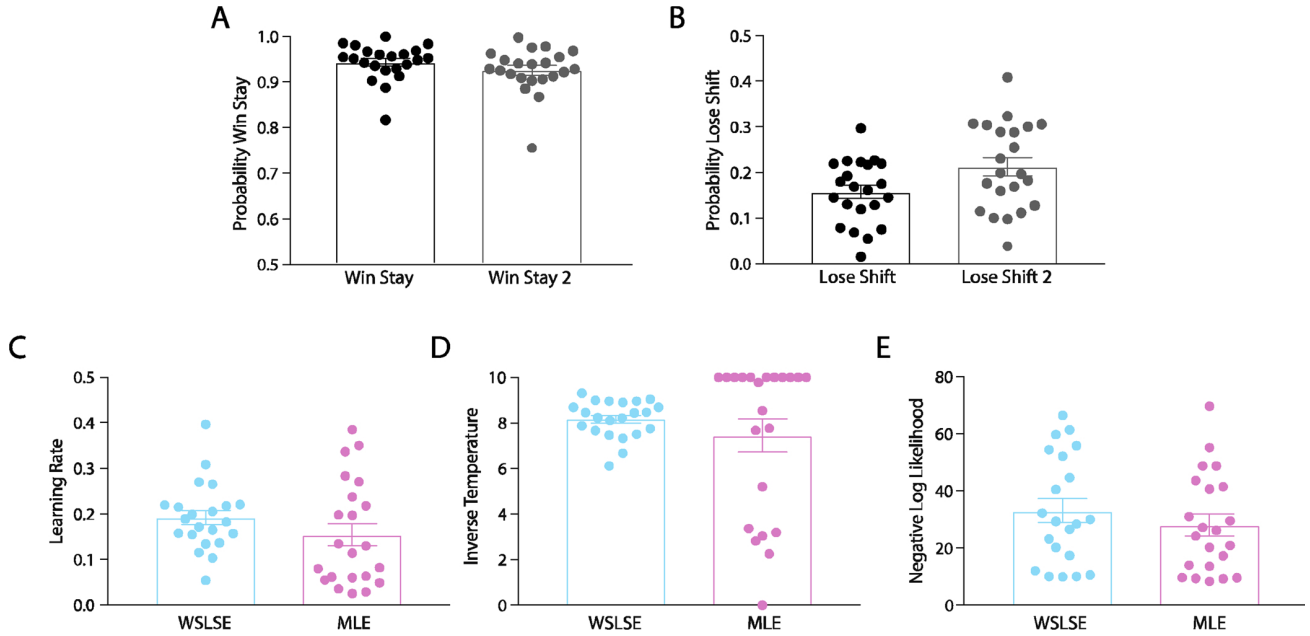

Fig. 7. (a) Win stay and win stay 2 behavior in a binary probabilistic choice task. (b) Lose shift and lose shift 2 behavior in a binary probabilistic choice task. All error bars indicate standard error of the mean. (c) Comparison of learning rate approximation using WSLSE and MLE in a binary probabilistic choice task (*t*-test comparison of means: t = 1.316, p = 0.1954; F-test for equality of variances: $F_{21,21}$ = 2.427, p = 0.0481). (d) Comparison of inverse temperature approximation using WSLSE and MLE in a binary probabilistic choice task (*t*-test comparison of means: t = 0.9828, p = 0.3313; F-test for equality of variances: $F_{21,21}$ = 17.41, p = 0.0001). (e) Comparison of negative log likelihood values calculated from the RL parameters approximated using WSLSE and MLE in a binary probabilistic choice task. All error bars indicate standard error of the mean.

RL parameter estimation. We found that our novel method of RL parameter estimation, WSLSE, performs with precision comparable to the conventional MLE approach under ideal conditions, as well as conditions with simulated random noise, and, critically, with empirical data with unknown noise. A major advantage of our model is the ease of implementation. Instead of performing maximum likelihood estimation across trial-by-trial data, our model simply makes use of the set of WSLS tendencies in a given learning task. This simplicity has the potential to increase the accessibility of RL parameter estimation, in turn fostering more translational approaches between the fields of behavioral and computational neuroscience.

Traditionally, higher win stay probabilities are interpreted as increased sensitivity to positive feedback while higher lose shift scores are interpreted as increased sensitivity to negative feedback (Bari et al., 2010; Paulus et al., 2002; St Onge et al., 2011). For the win stay metric, this means that the more individuals learn from positive prediction errors, the more likely they are to stay following a rewarding outcome. We indeed observe this relationship, with win stay probabilities increasing with learning rate, suggesting that the win stay metric corresponds to increased sensitivity to positive feedback. For lose shift, the traditional interpretation implies that the more individuals learn from negative prediction errors, the more likely they are to shift following a loss. Following this logic, we would expect a higher learning to correspond to higher lose shift probabilities, similar to the relationship between learning rates and win stay. What we see instead is a more parabolic relationship with both low and high learning rates corresponding to higher lose shift probabilities and more medial learning rates corresponding to lower lose shift probabilities. This more complex relationship is due to the fact that in order for individuals to be sensitive to loss, they must first have learned about reward. At low learning rates, this renders the lose shift metric meaningless. This challenges simplistic interpretations of the lose shift metric. Additionally, because of the nature of the relationship between lose shift probabilities and learning rates, lose shift probabilities at both low and high learning rates look similar, further complicating any meaningful interpretation of lose shift probabilities without additional information about learning rates.

Our novel method of WSLSE of RL parameters essentially provides a means of translating between WSLS and RL frameworks. Though they share a similar logic, i.e. that the outcome of previous choice impacts current choice, these two frameworks have existed in parallel to one other with little cross-talk. Behavioral analysis within a WSLS framework has a long history in the field of behavioral neuroscience (Schusterman, 1962; Slotnick and Katz, 1974; Williams, 1972). Our method provides a unique opportunity to stimulate cross-talk between WSLS and RL-based analyses of behavior, which are increasingly used, by facilitating simple translation of published WSLS behavioral data into an RL context, shedding light on previously obscured exploitation parameters and learning rates to unify these literatures.

Our model makes use of a simple reinforcement learning agent. In recent years there has been increasing interest in more complex reinforcement learning models that incorporate features such as dual learning rates and stickiness parameters, (Balcarras et al., 2016; Gershman, 2015; Langdon et al., 2019; Noworyta-Sokolowska et al., 2019). Our WSLSE approach is readily adaptable to approximate RL parameters for more elaborate models; however, such modifications would require certain considerations. With increasing numbers of distinct parameters, potential parameter combinations increase exponentially, making the requisite simulations for WSLSE substantially more computationally intensive. However, these simulations need be performed once only. Of greater concern is that, with increasing parameters, the ability of a limited set of WSLS probabilities to predict said parameters becomes less reliable. WSLSE for models with a high number of free parameters will likely approximate parameters with extremely high standard deviations. This could be mitigated by incorporating additional task-specific behavioral metrics such as total rewards earned, number of reversals completed, or n-back trial behaviors in addition to WSLS probabilities, to further restrict the parameter space, and thereby increase parameter recoverability.

Models are powerful in their ability to describe behavior in terms of its discrete latent underlying processes. However, when different

models describe similar behaviors, it can become difficult to converse across models. In these situations, the approach that is most often taken is to either create a hybrid model that combines facets of both models, creating a single encompassing model or to directly compare the models to each other to try and identify the superior model. We have instead used one model (RL), to explain the behavior of another model (WSLS) and shown that the second model (WSLS) can also be used to approximate the processes underlying the first (RL). This suggests that, even though WSLS and RL describe slightly different behavioral processes, ultimately, they explain much of the same variance. In explicating this link, we have created a simple and flexible method of performing RL parameter approximation, which instead of using trial-by-trial data, simply uses bulk WSLS tendencies. This simplified approach, reduces barriers to implementing computational modeling of RL and will enable increased communication between computational and behavioral neuroscience, ultimately leading to both a proliferation of new research and integration of the existing wealth of published behavioral data into reinforcement learning.

CRediT authorship contribution statement

Eshaan S. Iyer: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Megan A. Kairiss:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - review & editing, Visualization. **Adrian Liu:** Methodology, Writing - review & editing. **A. Ross Otto:** Methodology, Writing - review & editing, Supervision. **Rosemary C. Bagot:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

This research was supported by an NSERC Discovery Grant, an NSERC Accelerator Supplement and an FRQNT Nouveaux Chercheurs award to RCB, and an NSERC Discovery Grant, NSERC Discovery Launch Supplement and CIFAR Azrieli Global Scholars Award to AL.

References

Ahn, W.Y., Vasilev, G., Lee, S.H., Busemeyer, J.R., Kruschke, J.K., Bechara, A., Vassileva, J., 2014. Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. Front. Psychol. 5, 849. https://doi.org/10.3389/fpsyg.2014.00849.

Balcarras, M., Ardid, S., Kaping, D., Everling, S., Womelsdorf, T., 2016. Attentional selection can be predicted by reinforcement learning of task-relevant stimulus features weighted by value-independent stickiness. J. Cogn. Neurosci. 28 (2), 333–349.

Ballard, I.C., McClure, S.M., 2019. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. J. Neurosci. Methods 317, 37–44. https://doi.org/10.1016/j.jneumeth.2019.01.006.

Bari, A., Theobald, D.E., Caprioli, D., Mar, A.C., Aidoo-Micah, A., Dalley, J.W., Robbins, T.W., 2010. Serotonin modulates sensitivity to reward and negative feedback in a probabilistic reversal learning task in rats. Neuropsychopharmacology 35 (6), 1290.

Bathellier, B., Tee, S.P., Hrovat, C., Rumpel, S., 2013. A multiplicative reinforcement learning model capturing learning dynamics and interindividual variability in mice. Proc Natl Acad Sci U S A 110 (49), 19950–19955. https://doi.org/10.1073/pnas.1312125110.

Dalton, G.L., Phillips, A.G., Floresco, S.B., 2014. Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. J. Neurosci. 34 (13), 4618–4626. https://doi.org/10.1523/JNEUROSCI.5058-13.2014.

Daw, N.D., 2011. Trial-by-trial data analysis using computational models. Decis. Mak. Affect Learn. Atten. Perform. XXIII 23 (1).

Gershman, S.J., 2015. Do learning rates adapt to the distribution of rewards? Psychon. Bull. Rev. 22 (5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3.

Gustafson, N.J., Daw, N.D., 2011. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. PLoS Comput. Biol. 7 (10), e1002235. https://doi.org/10.1371/journal.pcbi.1002235.

Herrnstein, R.J., 2000. The Matching Law: Papers in Psychology and Economics. Harvard University Press.

Kuchibhotla, K.V., Hindmarsh Sten, T., Papadoyannis, E.S., Elnozahy, S., Fogelson, K.A., Kumar, R., Froemke, R.C., 2019. Dissociating task acquisition from expression during learning reveals latent knowledge. Nat. Commun. 10 (1), 2151. https://doi.org/10.1038/s41467-019-10089-0.

Langdon, A.J., Hathaway, B.A., Zorowitz, S., Harris, C.B.W., Winstanley, C.A., 2019. Relative insensitivity to time-out punishments induced by win-paired cues in a rat gambling task. Psychopharmacology (Berl.). https://doi.org/10.1007/s00213-019-05308-x.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7 (4), 308–313.

Noworyta-Sokolowska, K., Kozub, A., Jablonska, J., Parkitna, J.R., Drozd, R., Rygula, R., 2019. Sensitivity to negative and positive feedback as a stable and enduring behavioural trait in rats. Psychopharmacology (Berl.). https://doi.org/10.1007/s00213-019-05333-w.

Otto, A.R., Taylor, E.G., Markman, A.B., 2011. There are at least two kinds of probability matching: evidence from a secondary task. Cognition 118 (2), 274–279.

Paulus, M.P., Hozack, N., Frank, L., Brown, G.G., 2002. Error rate and outcome predictability affect neural activation in prefrontal cortex and anterior cingulate during decision-making. NeuroImage 15 (4), 836–846.

Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Class. Cond. II Current Res. Theory 2, 64–99.

Schusterman, R.J., 1962. Transfer effects of successive discrimination-reversal training in chimpanzees. Science 137 (3528), 422–423.

Slotnick, B.M., Katz, H.M., 1974. Olfactory learning-set formation in rats. Science 185 (4153), 796–798.

St Onge, J.R., Abhari, H., Floresco, S.B., 2011. Dissociable contributions by prefrontal D1 and D2 receptors to risk-based decision making. J. Neurosci. 31 (23), 8625–8633. https://doi.org/10.1523/JNEUROSCI.1020-11.2011.

Stachenfeld, K.L., Botvinick, M.M., Gershman, S.J., 2017. The hippocampus as a predictive map. Nat. Neurosci. 20 (11), 1643–1653. https://doi.org/10.1038/nn.4650.

St-Amand, D., Sheldon, S., Otto, A.R., 2018. Modulating episodic memory alters risk preference during decision-making. J. Cogn. Neurosci. 30 (10), 1433–1441. https://doi.org/10.1162/jocn_a_01253.

Sutton, R.S., Barto, A.G., 2011. Reinforcement Learning: An Introduction.

Thorndike, E.L., 1927. The law of effect. Am. J. Psychol. 39 (1/4), 212–222.

Verharen, J.P., Adan, R.A., Vanderschuren, L.J., 2019. Differential contributions of striatal dopamine D1 and D2 receptors to component processes of value-based decision making. Neuropsychopharmacology 1–10.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Bright, J., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 1–12.

Watkins, C.J.C.H., 1989. Learning from Delayed Rewards.

Wilhelm, S., Manjunath, B., 2010. tmvtnorm: A package for the truncated multivariate normal distribution. Sigma 2 (2).

Williams, B.A., 1972. Probability learning as a function of momentary reinforcement probability 1. J. Exp. Anal. Behav. 17 (3), 363–368.

Wilson, R., Collins, A., 2019. Ten simple rules for the computational modeling of behavioral data. Elife 8, e49547. https://doi.org/10.7554/eLife.49547.

Wimmer, G.E., Braun, E.K., Daw, N.D., Shohamy, D., 2014. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. J. Neurosci. 34 (45), 14901–14912. https://doi.org/10.1523/JNEUROSCI.0204-14.2014.

Worthy, D.A., Maddox, W.T., 2012. Age-based differences in strategy use in choice tasks. Front. Neurosci. 5, 145. https://doi.org/10.3389/fnins.2011.00145.

Worthy, D.A., Maddox, W.T., 2014. A comparison model of reinforcement-learning and win-stay-Lose-Shift decision-making processes: a tribute to W.K. Estes. J. Math. Psychol. 59, 41–49. https://doi.org/10.1016/j.jmp.2013.10.001.

Wunderlich, K., Rangel, A., O'Doherty, J.P., 2009. Neural computations underlying action-based decision making in the human brain. PNAS 106 (40).