Joint Modeling of Reaction Times and Choice Improves Parameter Identifiability in Reinforcement Learning Models

Ian C. Ballard[1,2], Samuel M. McClure[2]

1. Neurosciences Graduate Training Program, Stanford University. Stanford, CA 94305, USA.
2. Helen Wills Neuroscience Institute, University of California, Berkeley. Berkeley, CA, 94720, USA.
3. Department of Psychology, Arizona State University, Tempe, AZ 85287, USA.

Corresponding Author: Ian Ballard, iancballard@gmail.com. 10 Giannini Hall, Berkeley CA, 94720.

Declarations of Interest: None.

Author Contributions: IB conceptualized the study, conducted the analysis, and wrote the manuscript. SM provided supervision and critical revisions.

Keywords: Q-learning, delay discounting, intertemporal choice, parameter estimation, Bayesian statistics, reproducibility

Abstract

Reinforcement learning models provide excellent descriptions of learning in a variety of tasks. Many researchers are interested in relating parameters of reinforcement learning models to psychological or neural variables of interest. We demonstrate that parameter identification is difficult due to the fact that a range of parameter values provide approximately equal quality fits to data. This identification problem has a large impact on power: we show that a researcher who wants to detect a medium sized correlation (r = .3) with 80% power between a psychological/neural variable and learning rate must collect 60% more subjects than specified by a typical power analysis in order to account for the noise introduced by model fitting. We introduce a method that exploits the information contained in reaction times to constrain model fitting and show using simulation and empirical data that it improves the ability to recover learning rates.

1 Introduction

Rescorla and Wagner first specified how animal learning could be understood using a computational model in which learning is driven by the difference between expectations and outcomes (Rescorla & Wagner, 1972). In the nearly fifty years since this seminal work, there has been an explosion of interest in using computational reinforcement learning (RL) models to understand behavior and to characterize the functions of neural systems (Niv, 2009). These models are parameterized by the learning rate, which controls the relative weighting of recent versus older information in generating predictions. The learning rate is of considerable experimental interest as a dependent variable in experiments that influence learning (Behrens, Woolrich, Walton, & Rushworth, 2007), or as a means to understand inter-individual variability (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007) or neural function (Gläscher & O'Doherty, 2010; Schönberg, Daw, Joel, & O'Doherty, 2007). However, the learning rate and other parameters of reinforcement learning models are difficult to faithfully estimate (S. Gershman, 2016), increasing probability of both type I and type II errors. Methods for improving the reliability of these estimates would increase the utility of applying reinforcement learning models to the study of behavior, neural data, and disease (Maia & Frank, 2011).

Parameters of RL models are difficult to estimate primarily because there is a tradeoff between the learning rate and a second parameter, decision noise, which specifies how noisily subjects choose the higher-valued option. Any sequence of choices is roughly equally well described as having been produced by an agent who learns quickly but decides noisily or a subject who learns more slowly but decides more deterministically. This relationship is exacerbated by the fact that changes in the learning rate only modestly change predictions about choice. Our approach harnesses reaction times to help constrain estimates of learning rate and reduce this tradeoff between parameters. Variability in reaction times provides insight into hidden psychological variables such as how subjects integrate information, are influenced by frames or context, or navigate a speed/accuracy tradeoff (Ratcliff & McKoon, 2008; Stone, 1960). Further, reaction times have been used to fit a reinforcement learning model in a task without value-based choice (Bornstein & Daw, 2012). Because subjects should respond more slowly when the values of the options are more similar, reaction times provide information about the values learned by the subject and constrain the estimate of parameters that govern the

learning of those values. We derive a method for optimally weighting predictions from choice and reaction times in order to fit reinforcement learning models.

A second reason that parameters are difficult to estimate is that there are constraints on the amount of data that can be reasonably collected from an individual subject. When experimenters fit flexible models such as RL to limited data, they tend to overfit to noise in the data. The use of Bayesian priors can prevent overfitting by rendering certain parameter values unlikely, which reduces the effective complexity of the model and also reduces the tradeoff between correlated parameters (S. Gershman, 2016). We compare our reaction time method to the use of Bayesian priors and assess whether they can have an additive effect on the improvement of parameter identifiability.

2 Materials and Methods

We contrast parameter estimation in two tasks: a reinforcement learning task and a delay discounting task. Code for all simulations and analyses can be found at https://github.com/iancballard/RL-Tutorials.

2.1 Task specification: reinforcement learning

We consider a 2-armed bandit task in which the agent decides between two options that independently vary in their probability of reward. Bandits were initialized with a bad arm (35% chance of reward) and a good arm (65% chance of reward). On each trial, the probability of reward for each arm was updated as independent random walks. Thus, the probability of reward on each arm was changed by a random amount taken from Gaussian distribution, $N(0,.025)$, with reflecting upper and lower boundaries at 75% and 25% reward, respectively. This design is used to encourage learning over the course of many trials (Daw, Gershman, Seymour, Dayan, & Dolan, 2011).

We simulated trajectories of a reinforcement learning agent through the task. The agent tracked values V for each of the bandits s_i. On each trial, the agent updates the value of the chosen bandit:

$$V(s_{chosen})^{t+1} = V(s_{chosen})^{t} + \alpha\delta^{t} \quad (1)$$

where α is the learning rate and δ^t is the prediction error from trial t:

$$\delta^{t} = r^{t} - V(s_{chosen})^{t} \quad (2)$$

and r^t is the reward received on trial *t*. Values (*V*) were initialized at .5, as there are no negative rewards in this task and this allows for symmetric learning about reward and no-reward outcomes early in the task. Values were transformed into choice probabilities according to a softmax decision rule. If c^t is the choice on trial *t*,

$$p(c^t = s_i) = \frac{e^{mV(s_i)^t}}{\sum_j e^{mV(s_j)^t}} \quad (3)$$

where *m* is the inverse temperature parameter controlling choice variability.

2.2 Task specification: delay discounting

Delay discounting requires people to choose between rewards of different amounts (a_1, a_2) available after different delays (d_1, d_2) from the time of the experiment. Hyperbolic discounting describes the subjective value of rewards based on amount and delay:

$$V(a, d) = \frac{a}{1+kd} \quad (4)$$

where *k* is the individually determined discount rate that increases with preference for near-term rewards over larger, delayed rewards. To create a choice set, we sampled trial-wise *k*s from the same distribution used to generate subject *k*s (see 2.2) and, for each *k*, we set (a_1, a_2, d_1, d_2) to values such that an imaginary subject with that *k* would be indifferent between the two options. This procedure ensures that the choice-set spans the range of indifference points present in the simulated subject population.

Choices in delay discounting are assumed to be described by the softmax function specified in Eqn. 3 except that $V(a_1,d_1)$ and $V(a_2, d_2)$ replace $V(s_j)$. Two parameters therefore determine individual choice behavior in delay discounting: the discount rate, *k*, and the inverse temperature, *m*.

2.2 Simulation procedures

Because researchers are often interested in relating learning rates and discount rates to psychological or neural variables of interest, we assessed the extent to which it was possible to recover known parameters from a population of simulated subjects. For each simulation, we drew parameter settings for each subject and generated behavior using the reinforcement learning and delay discounting agents described above. We then fit the parameters of agents to the

synthetic data using the Scipy minimize function. We fit the data using four different approaches:

1) Maximum likelihood (ML). This standard technique maximizes the likelihood of the synthetic choice data.
2) ML with reaction times. This technique jointly maximizes the likelihood of the choice and reaction time data.
3) Maximum a posteriori (MAP). This technique uses Bayesian priors to maximize the posterior probability of the choice data.
4) MAP with reaction times. This technique maximizes the posterior probability of the choice and reaction time data.

Finally, we assessed the fitted parameter estimates against the ground truth parameters using Pearson correlation.

We assessed correlations between actual and fitted parameters for each modeling technique and for different numbers of subjects and bandit trials. For set of parameter values, we ran 1,000 simulations. We drew parameters from distributions that matched the expected distribution expected in the population, based on previous literature (Figure S1a, (Daw et al., 2011). A learning rate equal to 0 indicates no learning, whereas a learning rate of 1 indicates a win-stay loose shift strategy. However, low α does not necessarily indicate poor performance; rather, it indicates a smaller recency bias in weighting information over past trials (Bayer & Glimcher, 2005). We therefore allowed α to span its entire range of values. For delay discounting, we sampled *log(k)* from $N(-4.5, .9)$, Figure S1b (Ballard et al., 2017). For both models, we allowed the inverse temperature to vary from variable ($m = .5$) to nearly deterministic ($m = 10$), Figure S1c. The lower bound on m excludes completely noisy subjects; purely random choices are modeled instead as subjects with no learning. Predictions of behavior with α=0, m>0 are equivalent to α>0, m=0, so we ignore the m=0 condition without loss of generality. Parameter priors were used to generate simulated subjects and were also used as the Bayesian priors in MAP fitting.

In order to estimate parameter correlations across a range of parameter settings to create Figure 1c,d, we generated simulated data for 1,000 subjects on 200 trials for each cell within a grid of parameter settings. For each simulation run, we numerically computed the Hessian matrix at the maximum likelihood estimate. This matrix describes the convexity of the likelihood

function at its peak. We then inverted this matrix to obtain an asymptotic estimate of the variance-covariance matrix. Finally, we converted this matrix to a correlation matrix in order to obtain estimates of the correlation between the parameters at the peak of the likelihood function. Figure 1c,d show the means of these correlations across simulation runs for each parameter setting.

2.2 Reaction time modeling

We sought to jointly model the probability of the reaction time (*rt*) and choices (*c*). For each trial *t*:

$$p(c^t = s_i, rt^t | \, c^1 \dots c^{t-1}, rt^1 \dots rt^{t-1}, r^1 \dots r^{t-1}) \quad (5)$$

where r^t is the reward earned on trial *t*. Using the chain rule for probabilities, we can rewrite this equation:

$$p(c^t = s_i, | \, c^1 \dots c^{t-1}, rt^1 \dots rt^{t-1}, r^1 \dots r^{t-1}) \, p(rt^t | \; c^t = s_i, c^1 \dots c^{t-1}, rt^1 \dots rt^{t-1}, r^1 \dots r^{t-1}). \quad (6)$$

We now make the simplifying assumption that choice on trial *t* is conditionally independent of reaction times on *previous* trials:

$$p(c^t = s_i, | \, c^1 \dots c^{t-1}, r^1 \dots r^{t-1}) \, p(rt^t | \; c^t = s_i, c^1 \dots c^{t-1}, rt^1 \dots rt^{t-1}, r^1 \dots r^{t-1}). \quad (7)$$

This assumption is sure to be partially invalid because previous trial reaction times are related to previous trial values, which in turn are indirectly related to current trial choice. The complexity of modeling this relationship simultaneously with reinforcement learning is beyond the scope of the current work.

In order to specify a conditional probability distribution on reaction times, we specify a linear regression model relating values of the *p* options as well as the choice to reaction times. We make use of the fact that linear regression can be equivalently specified as a maximum likelihood solution to a linear probabilistic generative model:

$$rt^t = \sum_k \beta_k f_k(V_1^t, \dots, V_p^t, c^t) + \varepsilon, \quad (8)$$

where $\varepsilon \sim N(0, \sigma^2)$. For *n* trials, one can show that the joint log-likelihood of both the observed choices and reaction times is equal to:

$$\sum_{t=1}^{n} \log \left(\frac{e^{mV(s_i)^t}}{\sum_j e^{mV(s_j)^t}}\right) + n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(rt^t - \sum_k \beta_k f_k(V_1^t, \dots, V_p^t, c^t))^2. \quad (9)$$

The first term is the softmax log likelihood of choices and the second term is the likelihood of a linear regression model of reaction times, which is included in the output of most linear regression software packages (e.g., statsmodels in Python or logLik in R).

For our simulation data, we created reaction times that were a function of the bandit values. Given the pervasive finding that reaction times are slower for more difficult decisions (Ratcliff & McKoon, 2008), we defined reaction times to be a function of the absolute value of the difference in values between the bandits:

$$rt^t = -\beta|V_1^t - V_2^t| + \varepsilon. \quad (10)$$

We set $\beta = 1$ and added Gaussian random noise to the simulated reaction times with mean of 0 and standard deviation equal to 5 times the standard deviation of $|V_1^t - V_2^t|$ across trials. This procedure resulted in noisy reaction times that were correlated with the absolute value in the difference in choice options with an average $R^2 = .037$. We used a simple model of RTs for the sake of simplicity in simulations. However, one can define a more realistic model and we do so in our analysis of real bandit data.

4 Results

We aim to identify and address a problem of identifying model parameters in RL. The problem derives from the fact that the primary model parameter of interest in RL, the learning rate, is mostly constrained by trial-to-trial changes in behavior. We show below that this constraint leads to redundancy with the inverse temperature parameter that is intended to model choice variability. The parameter identification problem that we highlight is specific to RL and is not a general problem of fitting decision models of behavior. We perform parallel analyses with delay discounting to highlight the conditions under which parameter confusability should be expected.

4.1 Parameter identification in reinforcement learning

Parameter identification in reinforcement learning models is encumbered by an inherent relationship between learning rate and decision noise. Imagine a two-armed bandit task in which you have observed a subject make a leftward choice and observe an outcome. According to the model, the value of the leftward choice changes according to

$$V(s_L)^{t+1} = V(s_L)^t + \alpha\delta^t. \quad (11)$$

In the simplified two-choice case that we are considering, the probability of leftward choice is

$$p(c^t = s_L) = \left[1 + e^{-m\left(V(s_L)^t - V(s_R)^t\right)}\right]^{-1} \quad (12)$$

so that the probability of a leftward choice on the next trial is

$$p(c^{t+1} = s_L) = \left[1 + e^{-m\left(V(s_L)^t + \alpha\delta^t - V(s_R)^t\right)}\right]^{-1} \quad , \quad (13)$$

which can be rewritten as

$$p(c^{t+1} = s_L) = \left[1 + e^{-m\alpha\delta^t} e^{-m\left(V(s_L)^t - V(s_R)^t\right)}\right]^{-1}. \quad (14)$$

The critical thing to notice is that trial-to-trial changes in behavior (equation 12 compared with equation 14) depend on the prediction error (δ^t) and the product of $m \times \alpha$. A high precision (large m), slow learning (small α) subject would therefore make a similar proportional change in behavior on the subsequent trial as a low precision (small m), fast learning (large α) subject.

The degree to which learning rate and decision noise are exchangeable presents a problem because it means that some non-unique set of parameter values provides roughly equally good accounts of choice data. This problem is evident in model simulations. Figure 1A shows the likelihood surface for a reinforcement learning model of a simulated subject (α = .3, m = 2) in a two-arm bandit task. The dark blue area shows a strong tradeoff between parameters; a lower learning rate and reliable responding or higher learning rate and noisier responding provide similar accounts of the data. Figure 1C shows that this tradeoff exists over a wide range of parameters: the average absolute correlation between α and m at the peak of the likelihood surface is .49. One consequence of this type of likelihood surface is that the maximum likelihood estimate (the peak of the surface in 2A) can move large distances in parameter space with small changes in the data. That is, if this subject had chosen only a little differently, there could be a large change in the maximum likelihood estimate of learning rate.

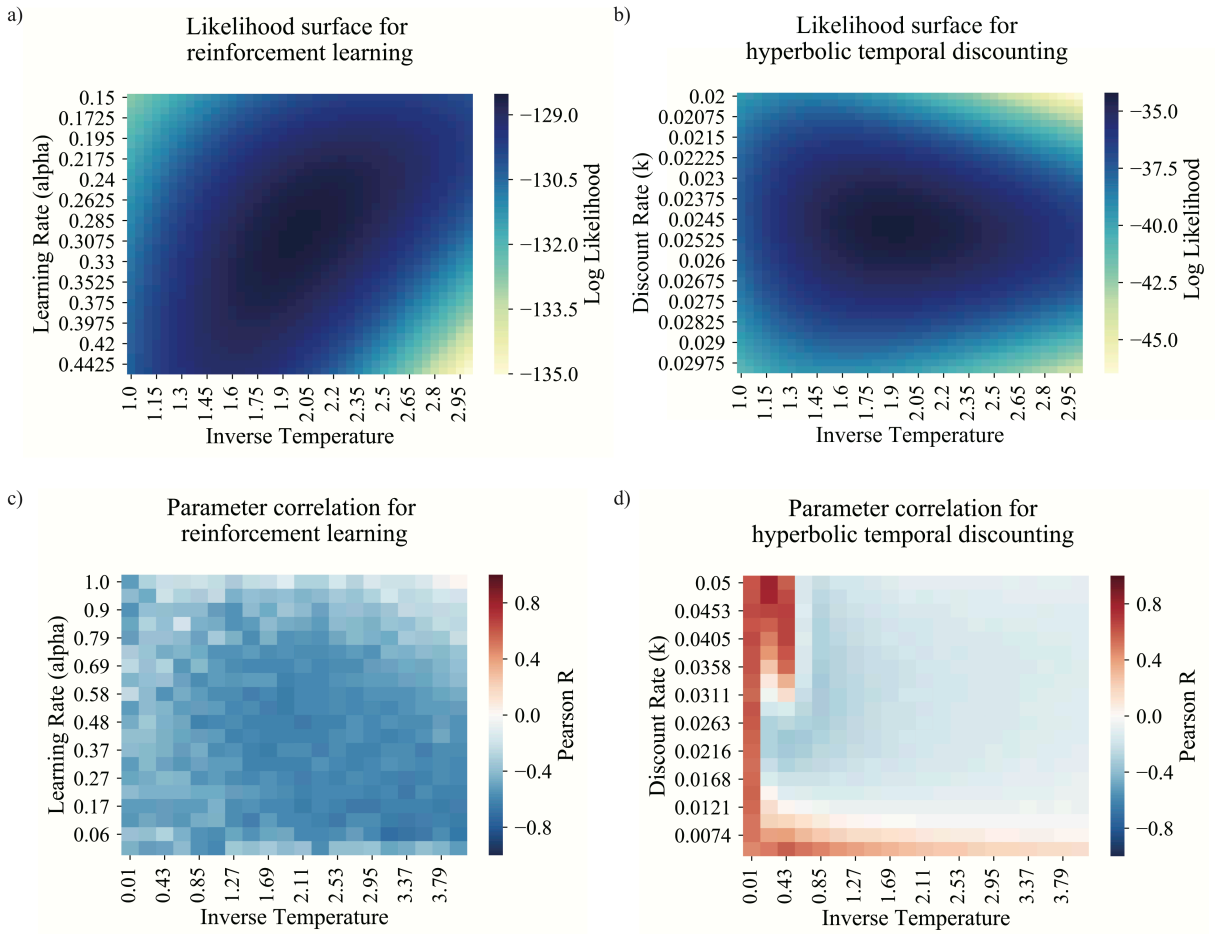

Figure 1. Likelihood surface of reinforcement and delay discounting models. A) Likelihood surface for a reinforcement learning model of a simulated subject on a 2-arm bandit task (α = .3, m = 2). There is a tradeoff between learning rate and inverse temperature, such that a lower learning rate and more reliable responding provides a similar fit as a higher learning rate and more random responding. B) Likelihood surface for a hyperbolic model of a simulated subject in an delay discounting task (k = .025, m = 2). Most of the uncertainty comes from the inverse temperature parameter. Compared to A, there is only a modest tradeoff between discount rate and choice noise. C) Estimates of the parameter correlation between α and m at the maximum likelihood estimate for data simulated from a range of parameter settings. Parameter anticorrelation is high for nearly the full range of parameter settings. D) Estimates of the parameter correlation between k and m at the maximum likelihood estimate for data simulated from a range of parameter settings. Parameter correlation is high for very noisy subjects and values of k near the edges of the choice set, but is generally low for a wide range of typical parameter settings.

In principle, the inherent relationship between learning rate and response variability should be ameliorated by the fact that value in RL models depends on all past outcomes. If $V(s_i)^0$ is the initial value for action s_i then

$$V(s_i)^n = \sum_{k=1}^{n} \alpha(1-\alpha)^{n-k} r^k + (1-\alpha)^n V(s_i)^0 \quad (15)$$

where n indexes trials in which s^i is chosen (and hence differs from t, used above). For all learning rates greater than zero, the most recent trial has the strongest effects on choice with an exponentially decreasing dependence on prior trials (i.e. $(1-\alpha)^{n-k}$). As a result, the differences in predictions between models with different learning rates can be subtle. They differ in the extent to which distant trials influence choice, but these distant trials always exert a much smaller effect than the most recent trial. The weak dependence on prior trials relative to the most recent trial is borne out in Fig. 1A,C: learning rate and temperature are not perfectly substitutable due to the effect of previous trials on estimated α, but there is a strong correlation between α and m in the log-likelihood isosurfaces.

4.2 Parameter estimation in delay discounting

Decision-making noise does not strongly interact with valuation parameters in decision tasks generally. As counterpoint to RL, we also explored parameter estimation in delay discounting. For modeling purposes, we assumed that people discount reward for delay according to the hyperbolic discounting model (Eqn. 4) and that choices are made using the softmax decision function (Eqn. 3). We studied parameter fitting for a binary choice task in which outcomes consist of a sooner reward, of amount a_s available at delay d_s, and a later reward, or amount a_ℓ and delay d_ℓ. The probability of choosing the smaller reward is therefore

$$p(c^t = \{a_s, d_s\}) = \left[1 + e^{-m(V_s - V_\ell)}\right]^{-1}. \quad (16)$$

The exponent in this equation is equal to

$$-m(V_s - V_\ell) = -\frac{ma_s}{1+kd_s} - \frac{ma_\ell}{1+kd_\ell}. \quad (17)$$

The relationship between the parameters of the model (k and m) and choice is a function of the amount and delay of the rewards, which change trial to trial, and the parameters are not generally exchangeable. However, when $kd_s \gg 1$ and $kd_\ell \gg 1$ then this equation is approximated by

$$-m(V_s - V_\ell) \approx -\frac{m}{k}\left(\frac{a_s}{d_s} + \frac{a_\ell}{d_\ell}\right) \quad . \quad (18)$$

In this regime, m and k are exchangeable. However, k is generally at most on the order of 10^{-2} (units of day^{-1}), intertemporal choice task typically include immediate rewards (d_s=0), and the longest delays are on the order of one year (10^2 days). For every delay discounting task that we are aware of, the approximation in Eqn. 18 will not hold for the vast majority of trials. It is

therefore generally not the case that *m* and *k* are interchangeable when fitting to delay discounting data. This is evident in Figure 1B. The inverse temperature is not well constrained in our modeled task, but there is relatively little correlation between *m* and *k* in log-likelihood isosurfaces. Over a range of parameter settings, the average absolute correlation between *m* and *k* is .19, and this correlation mostly arises at extreme values of the parameters, Figure 1D. This example illustrates that the problem we address is not general to psychological models with decision noise.

4.3 Reaction times and Bayesian priors improve parameter identifiability

Experiments typically aim to estimate RL learning rate in order to relate this variable to some other variable of interest (e.g., Behrens et al., 2007; Rouhani, Norman, & Niv, 2018). Errors in learning rate estimates introduce noise to the comparison and weaken the power of the subsequent test. We focused our analysis on our ability to recover learning rates from a cohort of subjects. For each cohort, we drew parameters for each subject, simulated a run through a two-arm bandit using the RL model described, and used the resulting data to estimate the generative model parameters. For each cohort, we computed the correlation between the ground truth learning rates and the recovered learning rates. We examined these correlations as a function of the number of trials of the task and the number of subjects in a cohort (Figure 2). As anticipated by the previous analyses, the ability to reconstruct discount rates in delay discounting is near ceiling (Figure S2). We therefore focus our analyses on parameter identifiability in RL.

We find that the ability to reconstruct learning rates is modest. For 100 trials, $r = .66$, and for 200 trials, $r = .79$. This number of trials is typical in bandit tasks (e.g., Wimmer, Braun, Daw, & Shohamy, 2014; Wunderlich, Rangel, & O'Doherty, 2009). Note that these simulations are likely to overestimate the true ability to reconstruct learning rates because the simulated subjects use an identical generative model to that used for fitting. It is likely that models used for fitting real data approximate all of the psychological processes at play during task performance. These correlation estimates can therefore be viewed as an approximate upper bound on the ability to fit learning rates. Increasing the number of trials improves the ability to reconstruct learning rates, but there are often financial and psychological limits to the maximum number of trials. Perhaps surprisingly, increasing the number of subjects does not appreciably improve the ability to reconstruct learning rates.

We next contrasted methods for improving parameter identifiability (Figure 2B). All tests were Bonferroni corrected for 10 tests across the trial and subject bins. Using either reaction times or Bayesian priors improved parameter identifiability for all trial counts and all cohort sizes, all $p < .001$. Reaction times improved parameter identifiability more than MAP estimation with Bayesian priors, except for bandits with 100 trials, $p > .2$, 15 subjects, $p = .019$, all others, $p < .001$. This finding is remarkable given that reaction times were noisily related to the difference in option values (average $R^2 = .037$) and simulated subject parameters were drawn from these prior distributions. Finally, the combined use of Bayesian priors and reaction times provided the best fit for all cohort sizes and numbers of trials, all $p < .001$. This finding confirms that reaction times and Bayesian priors offer partially distinct ways to regularize RL parameter estimates, and their combined use further improves estimates.

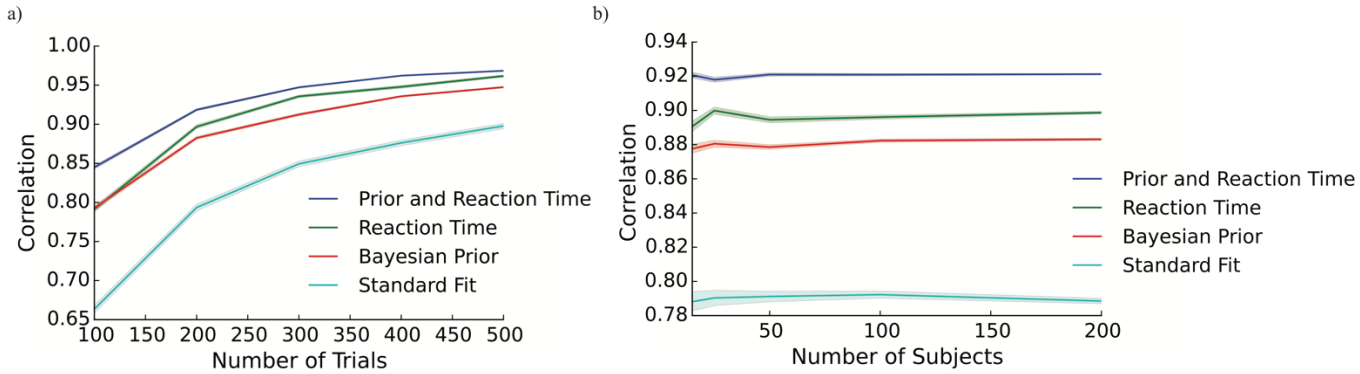

Figure 2 Simulation of parameter identifiability in RL. A) The correlation between ground truth and fitted learning rates as a function of the number of bandit trials (25 subjects). Increasing the number of trials improves parameter identifiability, and the use of reaction times and Bayesian priors substantially improves parameter identifiability regardless of the number of trials. B) The correlation between ground truth and fitted learning rates as a function of the number of subjects (200 trials). Increasing the number of subjects did not improve parameter identifiability. The use of reaction times and Bayesian priors substantially improves parameter identifiability regardless of the number of subjects.

4.4 Effect of parameter identifiability on experimental power

Our analysis of the ability to recover the learning rates of a subject population are optimistic. Even so, the simulated correlation between recovered learning rates and ground truth may appear high. However, even moderate decrements in the ability recover learning rates can

have a large effect on experimental power (Figure 3). For example, one needs around 85 subjects to detect a correlation of $r = .3$ with 80% power. If experimental constraints set the maximum number of trials to be 200, then our simulations suggest that experimenter would need to collect data from *at least* 137 subjects to detect the same effect with the same power, a 60% increase. According to our simulations, the use of reaction times could increase parameter identifiability sufficiently to require 105 subjects. The use of these model-fitting techniques may therefore have appreciable effects on the power.

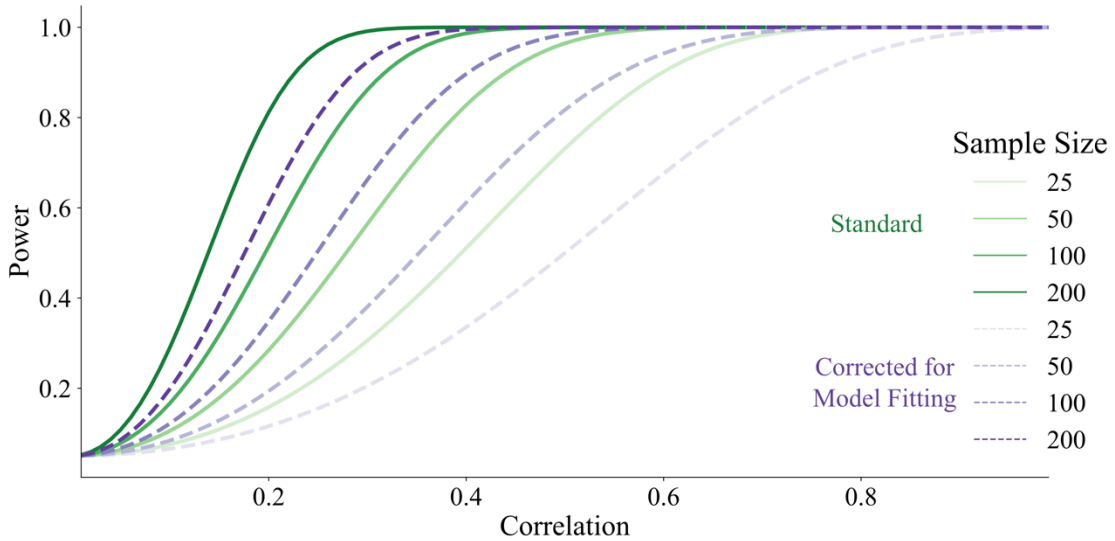

Figure 3. Effect of parameter identifiability on experimental power. Green lines depict the power to detect a correlation between two variables for different sample sizes. Purple lines depict the power after accounting for the noise introduced by model-fitting.

4.5 Application to Empirical Data

Our simulations showed that both the use of a Bayesian prior and the use of reaction times in model fitting improved parameter identifiability. Although these simulations were designed to match experimental data as closely as possible, they included two important features that could render their predictions over-optimistic. First, the parameters were generated from the same prior distribution used for fitting. Second, the reaction times were generated from the model used to fit reaction times. We therefore assessed the efficacy of these methods in a previously published dataset (Wimmer et al., 2014). Thirty subjects performed 100 trials of a two-armed bandit task in which each bandit was associated with a different, trial-unique object

that was irrelevant to the task. In addition, subjects performed a second run of the task without the objects. Although there were differences between the tasks (e.g., presence of objects), and there could have been state differences within a subject (e.g., fatigue), we reasoned that learning rates assessed from two very similar bandit tasks should be strongly correlated. We modeled log transformed reaction times as a function of 1) the linear and quadratic effects of the absolute value of the difference in values between the bandits 2) the linear effect of trial number, and an 3) indicator function on whether the subject choose the bandit with the maximum value:

$$log\ (rt)^t = \beta_o - \beta_1|V_1^t - V_2^t| - \beta_2|V_1^t - V_2^t|^2 - \beta_3 t - \beta_4 I(c^t = \max_i(V_i^t)) + \varepsilon. \qquad (10)$$

Where $\max_i(V_i^t)$ is the index of the maximum value and I is an indicator function.

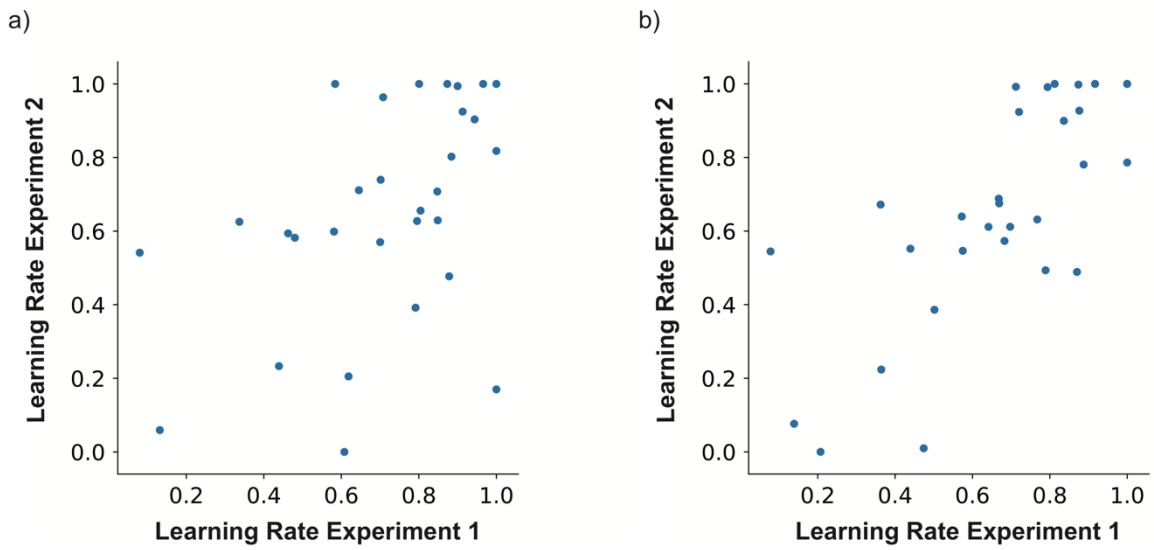

Figure 4. Joint modeling of choice and reaction times improves parameter identifiability in real data. A) The correlation in estimated learning rate between two runs of a bandit task using a standard RL model. B) The correlation in estimated learning rate when estimated using a model of reaction times and choice.

Under standard model fitting with no Bayesian priors and no reaction times, learning rates are correlated, $r(28) = .47$, $p = .009$, Figure 4A. Fitting with Bayesian priors results in a larger correlation, $r(28) = .59$. We used the R package Bayesian First Aid to assess the difference in correlations and found a 74.5% posterior probability that Bayesian priors improved the correlation. Fitting with reaction times results in an even higher correlation, $r(28) = .75$, Figure 4B. The Bayesian posterior probability that reaction times increased the magnitude of the correlation over standard model fitting is 94.5%. In contrast to our simulation results, the

combined use of both Bayesian priors and RT results in a similar strength correlation as using only Bayesian priors, $r(28) = .60$, $p < .001$. These parameters in turn were only moderately correlated with the parameters from using Bayesian priors, $r(28) = .62$, $p < .001$, or reaction times, $r(28) = .55$, $p = .002$, alone. We speculate that this is because our Bayesian priors do not perfectly align with the distribution of parameters in our subjects, and therefore reaction times and Bayesian priors have different effects on parameter estimates. Overall, our results suggests that fitting reaction times results in a higher correlation between learning rates assessed from the same subjects in different bandit tasks.

5 DISCUSSION

The utility of reinforcement learning models depends on the ability to estimate the parameters of the RL model. We showed that jointly fitting choices and reactions times improves the reliability of RL parameter estimates. Other promising approaches include the use of empirical priors derived from an independent, similar dataset (S. Gershman, 2016) or hierarchical fitting of group-level and subject-level parameters (Chávez, Villalobos, Baroja, & Bouzas, 2017). However, Bayesian priors actually interfered with the ability of reaction times to improve parameter identifiability in real data if the priors may have been misspecified (section 4.5). Fitting based on reaction times requires no prior knowledge of your subject or subject population. Moreover, the efficacy of the Bayesian prior approach is controversial (Spektor & Kellen, 2018) and it can yield counterintuitive results. For instance, the empirical prior on learning rate advocated by Gershman is concentrated entirely on 0 and 1. This prior expresses the belief that subjects do not use reinforcement learning: they either don't learn, or they do use a win-stay/lose-shift strategy. Although this prior is empirically derived, it expresses a belief about learning that is opposed by a wealth of behavioral and neural data. Nonetheless, the use of empirical priors is consistent with the approach advocated here, and we believe that the use of reaction time data could still benefit this approach.

Our modeling approach focused on a simple reinforcement learning model applied to a bandit task. There is substantial interest in relating parameters of more complex RL tasks to psychological variables. For example, in the two-step decision task, subjects make sequential decisions to earn rewards (Daw et al., 2011). This behavior is well-described by a hybrid

between a model-free learning agent, similar to the one described here, and a model-based agent that makes decisions based on a model of the sequential structure of the task. Experimenters have estimated the relative weighting between these two systems and have related the weighting parameter to individual differences in neural activation (Daw et al., 2011; Doll, Duncan, Simon, Shohamy, & Daw, 2015), working memory capacity (Otto, Raio, Chiang, Phelps, & Daw, 2013), habit persistence (Gillan, Otto, Phelps, & Daw, 2015) and compulsive behavior (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016). Because hybrid models depend on the same reinforcement learning mechanism described here, it is likely that at least some of the parameters in these models suffer from similar fitting problems. We expect that joint modeling of reaction times could help improve estimates of individual differences in goal-directed behavior.

Formal models are vital for developing a theoretical understanding of brain and behavior (Gläscher & O'Doherty, 2010). However, most models contain hidden complexity that can reduce their applicability. In RL, this complexity is due to the fact that the behavior of the model changes slowly with changes in parameterization (Wilson & Niv, 2015). Faced with this problem, experimenters should make use of all the information available to help constrain estimates of learning. Reaction times are a useful and readily available source of such information. Future work should consider how biometrics such as eye tracking could be used to further constrain model estimates (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017).

6 Acknowledgements

The authors would like to thank Elliott Wimmer for providing data, Yuan Chang Leong for feedback and the NSF GRFP and NSF IGERT grant for providing training support for I.B.

7 References

Ballard, I. C., Kim, B., Liatsis, A., Aydogan, G., Cohen, J. D., & McClure, S. M. (2017). More Is Meaningful: The Magnitude Effect in Intertemporal Choice Depends on Self-Control. *Psychological Science*. http://doi.org/10.1177/0956797617711455

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141. http://doi.org/10.1016/j.neuron.2005.05.020

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Publishing Group*, *10*(9), 1214. http://doi.org/10.1038/nn1954

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*(7), 1011–1023. http://doi.org/10.1111/j.1460-9568.2011.07920.x

Chávez, M. E., Villalobos, E., Baroja, J., & Bouzas, A. (2017). Hierarchical Bayesian modeling of intertemporal choice. *Judgement and Decision Making*, *12*(1), 19–28.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. http://doi.org/10.1016/j.neuron.2011.02.027

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767–772. http://doi.org/10.1038/nn.3981

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. http://doi.org/10.1073/pnas.0706111104

Gershman, S. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://doi.org/10.1016/j.jmp.2016.01.006

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*, e94778. http://doi.org/10.7554/eLife.11305

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523–536. http://doi.org/10.3758/s13415-015-0347-6

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 501–510. http://doi.org/10.1002/wcs.57

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. http://doi.org/10.1016/j.neuron.2016.12.040

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. http://doi.org/10.1038/nn.2723

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. http://doi.org/10.1016/j.jmp.2008.12.005

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 20941–20946. http://doi.org/10.1073/pnas.1312011110

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. http://doi.org/10.1162/neco.2008.12-06-420

Rescorla, R., & Wagner, A. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement (Appletone-Century-Crofts, New York).

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. http://doi.org/10.1037/xlm0000518

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *27*(47), 12860–12867. http://doi.org/10.1523/JNEUROSCI.2496-07.2007

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making : Empirical priors. *Psychonomic Bulletin & Review*, 1–22. http://doi.org/10.3758/s13423-018-1446-5

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. http://doi.org/10.1007/BF02289729

Wilson, R. C., & Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLoS Computational Biology*, *11*(6), e1004237–21. http://doi.org/10.1371/journal.pcbi.1004237

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *34*(45), 14901–14912. http://doi.org/10.1523/JNEUROSCI.0204-14.2014

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(40), 17199–17204. http://doi.org/10.1073/pnas.0901077106