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

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

1. Stanford Neurosciences Graduate Training Program, Stanford University. Stanford, CA 94305, USA
2. Department of Psychology, Arizona State University, Tempe, AZ 85287, USA.

Corresponding Author: Ian Ballard, iancballard@gmail.com. 450 Serra Mall, Stanford CA, 94305.

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, parameter estimation, Bayesian statistics, reproducibility

Abstract

Reinforcement learning models are excellent models of learning in a variety of tasks. Many researches are interested in relating parameters of reinforcement learning models to psychological or neural variables of interest. However, these parameters are difficult to estimate reliably because the predictions of the model about choice change slowly with changes in the parameters. This identifiability 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 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

In 1972, 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. This parameter 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 reinforcement learning models are difficult to estimate for several reasons. First among these is that there is a tradeoff between the learning rate and decision noise, which specifies how noisily subjects choose the higher-valued option. Any sequence of choices can be roughly equally described as an agent who learns quickly but decides noisily or a subject who learns slowly but decides more deterministically. Our approach is to harness reaction times to help constrain estimates of learning rate and reduce this tradeoff between parameters. Variability in reaction times can provide 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 can provide additional 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, resulting in parameter estimates that do not generalize. The use of Bayesian priors can prevent overfitting by rendering certain parameter 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

2.1 Task specification

Code for all simulations and analyses can be found at https://github.com/iancballard/RL-Tutorials. 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 independently from a Gaussian distribution with mean 0 and a standard deviation .025, with reflecting upper and lower boundaries at 75% and 25% reward, respectively. This type of 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:

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

Where α is the learning rate and δ^{t} is the prediction error:

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

and r^t is the reward received on trial t. Values 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,

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

Where m is the inverse temperature parameter controlling choice stochasticity.

2.2 Simulation procedures

Because researchers are often interested in relating learning rates to a psychological or neural variable of interest, we assessed the extent to which it was possible to recover known learnings rates from a population of simulated subjects. For each simulation run, we drew parameter settings for each subject and generated behavior using the reinforcement learning agent described above. We then fit the parameters of this agent to the synthetic data using the Scipy's 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 then assessed the fitted parameter estimates against the ground truth parameters using a Pearson correlation.

We assessed these correlations for each modeling technique and for different numbers of subjects and bandit trials. For each of these bins, we ran 1,000 simulations. We drew parameters from distributions that matched the expected distribution expected in the population, based on previous literature (Figure 1, (Daw et al., 2011). A learning rate equal to 0 indicates no learning, whereas a learning rate of equal to 1 indicates a win-stay loose shift strategy. However, low alpha does not necessarily indicate bad performance; rather, it indicates that subjects have a smaller recency bias in weighting information over past trials (Bayer & Glimcher, 2005). We therefore allowed alpha to span its entire range of values, with a median of .41. We allows the inverse temperature to vary from fairly stochastic ( ) to deterministic (, with the median of the distribution, = 2.78, corresponding to a choice rule that is roughly equivalent to probability matching. Our lower bound on *m* was meant to exclude completely noisy subjects because we allowed for subjects with no learning. Including too many of these subjects represents an overly pessimistic view of the subject population. These priors were used to generate simulated subjects and also as the Bayesian priors in MAP fitting.

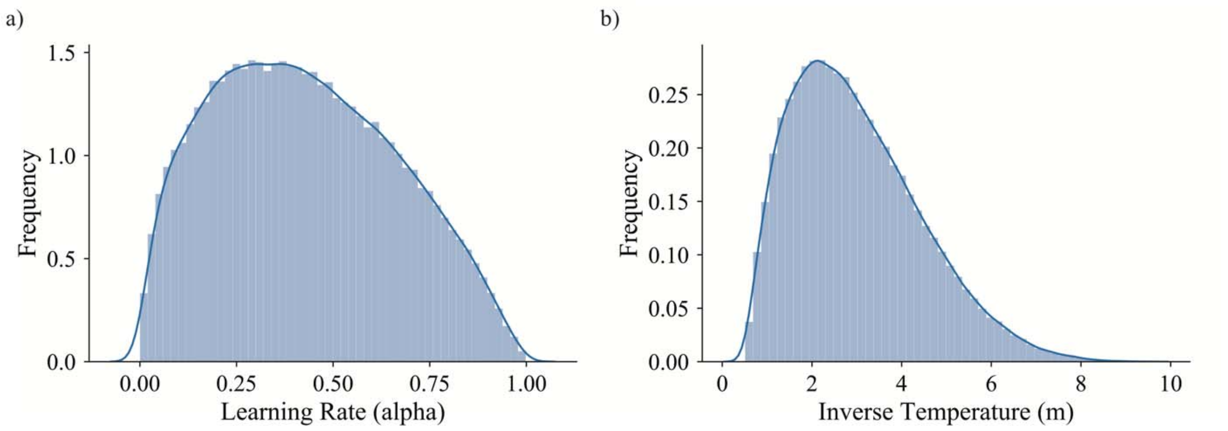

Figure 1 Prior distributions on RL parameters. A) Learning rates were drawn from a beta distribution with shape 1.5 and scale 2. B) Choice noise parameters were drawn from a beta distribution with shape 2 and scale 6, stretched by a factor of 10, and shifted to the right by .5.

2.2 Reaction time modeling

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

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

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

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

We now make the simplifying assumption that choices are conditionally independent of reaction times on *previous* trials:

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

This assumption is sure to be partially invalid, and future methods that explicitly model this relationship should perform better. However, 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:

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

Where $\varepsilon \sim N(0, \sigma^2)$.

For *n* trials, one can show that the log joint likelihood of both the observed choices and reaction times is equal to:

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

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 can be easily read from the output of most linear regression software packages (e.g., statsmodels in Python).

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:

$$rt^t = \beta|V_1^t - V_2^t| + \varepsilon$$

We set β = 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 the reaction time regressor. 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 complex and realistic model relating values and choices to reaction times and we do so in our analysis of real bandit data.

4 Results

4.1 Illustration of RL parameter identifiability

Parameter identifiability in reinforcement learning models is due to the relationship between learning rate and decision noise. Imagine a two-armed bandit task in which you have observed a subject make a leftward choice, receive a reward, and then subsequently make a rightward choice. It is impossible to determine whether the subject chose rightward on the second trial because she did not learn or because she learned but responded randomly. The strength of this correlation between learning rate and decision noise will depend on many factors, including the number of trials and the true learning rate and decision function. Nonetheless, any correlation between parameters presents a problem because it means that some span of parameter settings provide roughly equally good accounts of the data. Figure 1a shows the likelihood surface for a reinforcement learning model of a simulated subject in a two-arm bandit. The dark blue area shows a strong tradeoff between parameters; a lower learning rate and reliable responding or higher learning rate and more stochastic responding provide similar accounts of the data. 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.

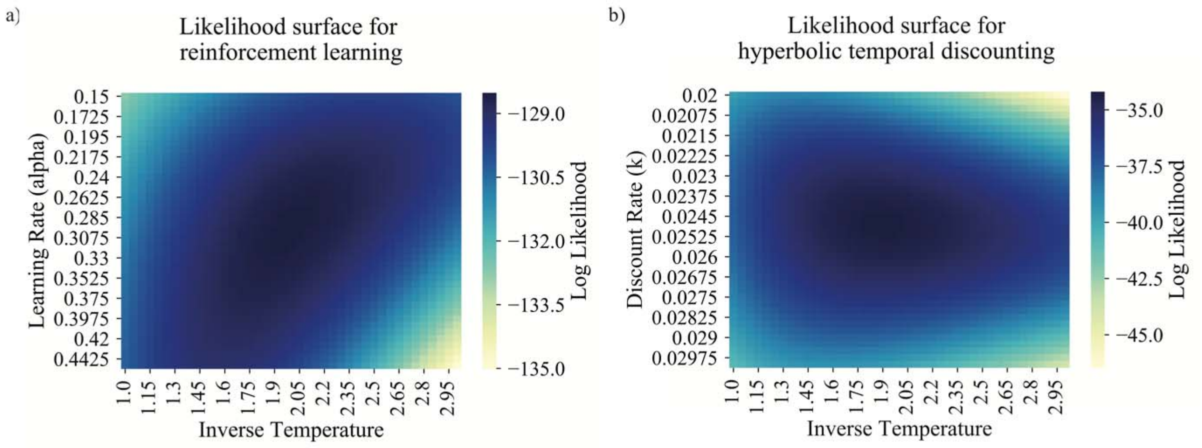

Figure 2. Likelihood surface of reinforcement and temporal discounting models. A) Likelihood surface for a reinforcement learning model of a simulated subject on a 2-arm bandit task (alpha = .3, inverse temperature = 2). There is a tradeoff between learning rate and inverse temperature, such that a lower 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 intertemporal choice task (discount rate = .025, inverse temperature = 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.

This tradeoff is a general property of psychological models that attempt to separately identify an underlying transformation of the experimental variables from a noisy decision function. However, RL models suffer from an additional problem that exacerbates the degree of correlation between learning rate and choice noise. The learning rate controls the relative weight of recent versus more distant trials in determining value. For all learning rates greater than 0, the most recent trials will have the strongest effects on choice. As a result, the differences in predictions between models with different learning rates can be subtle: they differ only in the extent to which distant trials influence choice, but these distant trials always exert a smaller effect than recent trials. Therefore, a range of learning rates can explain any given sequence of choices and rewards, exacerbating the correlation between learning rate and choice noise.

The property that choice predictions change slowly with changes in parameters is not a general property of psychological models. By way of illustration, consider a temporal discounting experiment in which subjects make choices between delayed rewards, (e.g., $5 now or $10 in 2 weeks). The subjective value of the rewards, discounted by the delay until their receipt, is typically modeled using a hyperbolic function:

$$Subjective\ Value = \frac{Amount}{1 + k * Delay}$$

And choices are typically modeled with the same softmax function used in RL. Imagine organizing the choices available to the subject by the how impatient they would need to be in order to be indifferent between the two options. A subject's *k* specifies where in this sorted list of choices she switches from choosing the shorter delay reward to the longer delay reward. As *k* changes, this switch point changes, making strong predictions about choice. As shown in Figure 1b, hyperbolic models have a weaker tradeoff between discount rate and decision noise, allowing for a more robust estimation of parameters than in reinforcement learning. Given the unique challenges of estimating the parameters of reinforcement learning models, we assess the utility of two complementary methods in constraining the values of these parameters.

4.2 Reaction times and Bayesian priors improve parameter identifiability

An experimenter is typically interested in the estimate of learning rate in order to relate this variable to some other variable of interest. Any error in estimation of learning rate will introduce noise to this comparison and weaken the power of the subsequent test. Therefore, 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 with a simple RL model, and used the resulting data to fit the parameters of the same model. 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.

We find that the ability to reconstruct learning rates is low for 100 trials, $r = .66$, or 200 trials, $r = .79$, experiments that are typical in bandit tasks (Figure 3). Note that these simulations are likely to overestimate the true ability to reconstruct learning rates because the simulated subjects use a generative model that is then used to fit behavior. These 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. While larger samples increase the power of the correlation test, it is also harder to discriminate between subjects as the average difference in learning rate between subjects decreases.

We next turned to analyzing our methods for improving parameter identifiability. All tests were Bonferroni corrected for 10 tests across the trial and subject bins. Using either reaction times or priors improves 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 the 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 the parameter estimates, and their combined use can potentially have an additive effect.

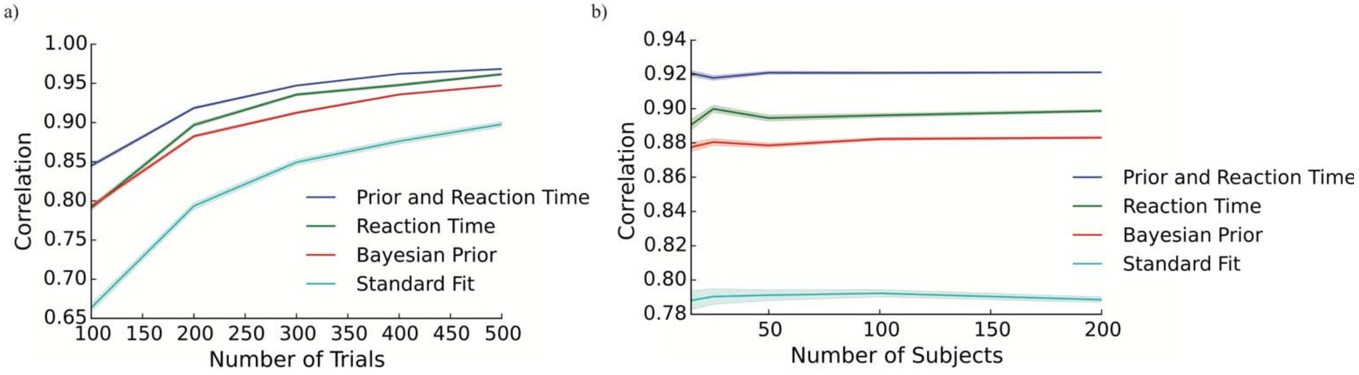

Figure 1 Simulation of parameter identifiability. A) The correlation between ground truth and fitted learning rates as a function of the number of bandit trials. Increasing the number of trials helps parameter identifiability, and the use of reaction times and Bayesian priors substantially improves parameter 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. 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.

4.3 Effect of parameter identifiability on experimental power

Our estimates 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 4). 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 can therefore have appreciable effects on the power.

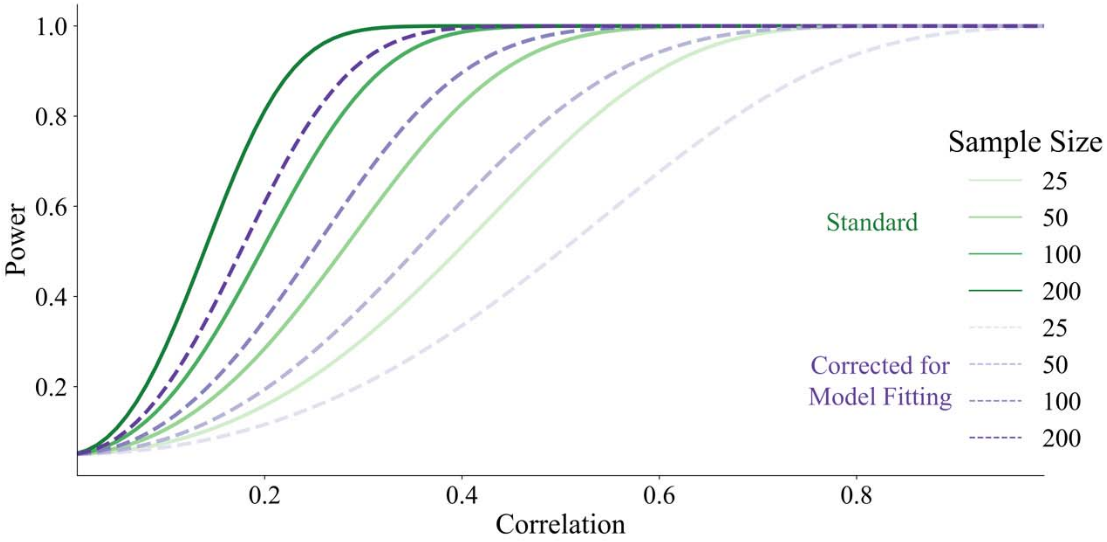

Figure 4. Effect of parameter identifiability on experimental power. Green lines depict the power to detect a correlation between two variables for different sample sizes. Purple lines depict the power after accounting for the noise introduced by model-fitting.

4.4 Application to Empirical Data

Simulations show that both the use of a Bayesian prior and the use of reaction times in model fitting helps 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, Braun, Daw, & Shohamy, 2014). 30 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 effect 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.

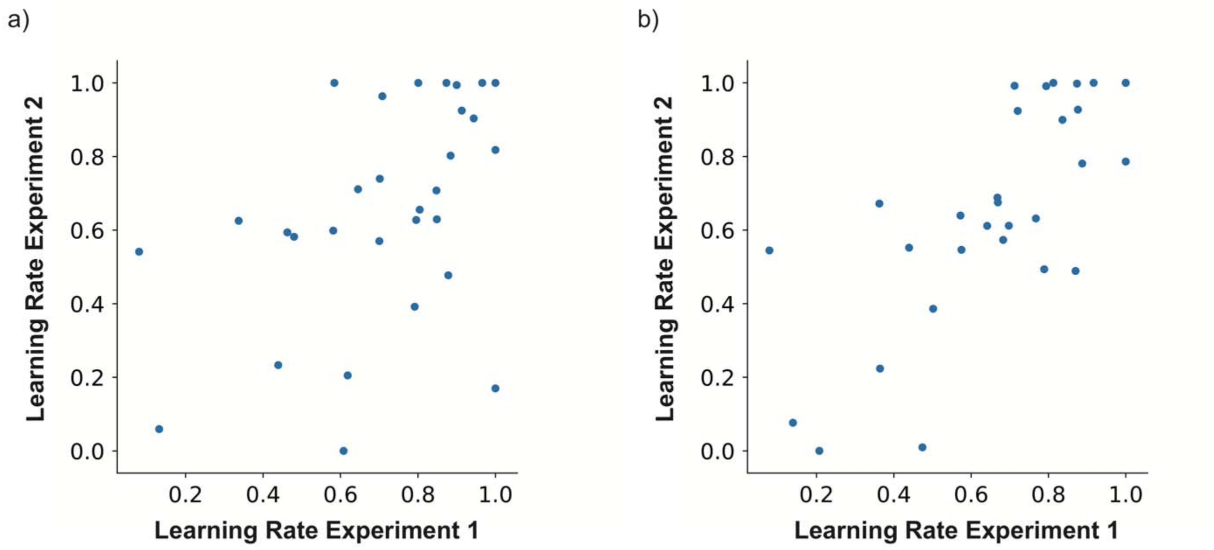

Figure 5. Joint modeling of choice and reaction times improves parameter identifiability in real data. A) The correlation in estimated learning rate between two runs of a bandit task using a standard RL model. B) The correlation in estimated learning rate when estimated using a model of reaction times and choice.

Under standard model fitting with no Bayesian priors and no reaction times, learning rates are correlated, *r*(28) = .47*, p* = .009, Figure 5a. 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 5b. 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 push parameter estimates in different directions. 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 for understanding individual variability or individual neural responses depends on the ability to estimate the parameters of the model. We showed that jointly fitting choices and reactions times or the use of Bayesian priors can improve the reliability of these parameter estimates. Another promising approach is to use empirical priors derived from an independent, similar dataset (S. Gershman, 2016) or a hierarchical approach that simultaneously fits group-level and subject-level parameters (Chávez, Villalobos, Baroja, & Bouzas, 2017). This approach is particularly promising given our finding that our Bayesian priors actually interfered the ability of reaction times to improve parameter identifiability in real data where the priors may have been misspecified. However, the efficacy of this 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 has the benefit of being empirically derived, the fact that it expresses a belief about learning that is opposed by a wealth of behavioral and neural data runs somewhat counter to the spirit of a Bayesian prior. Nonetheless, the use of empirical priors is consistent with the approach advocated here, and we believe that the use of reaction time data would 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 attempted to measure the relative weighting between these two systems and relate this 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 these models depend on the same reinforcement learning mechanism described here, it is likely that the weighting parameter is similarly difficult to estimate. We anticipate 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 and are practically useful tools for distilling the information contained in data (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

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. *Dx.Doi.org.Stanford.Idm.Oclc.org*, *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).

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