Do learning rates adapt to the distribution of rewards?

Citation: Gershman, Samuel J. "Do Learning Rates Adapt to the Distribution of Rewards?" Psychonomic Bulletin & Review 22.5 (2015): 1320–1327.

As Published: http://dx.doi.org/10.3758/s13423-014-0790-3

Publisher: Springer US

Persistent URL: http://hdl.handle.net/1721.1/103813

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Do Learning Rates Adapt to the Distribution of Rewards?

Samuel J. Gershman

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Word count: 3708

Address for correspondence:

Samuel Gershman

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

77 Massachusetts Ave., Room 46-4053

Cambridge, MA 02139

Phone: 773-607-9817

E-mail: sjgershm@mit.edu

Abstract

Studies of reinforcement learning have shown that humans learn differently in response to positive and negative reward prediction errors, a phenomenon that can be captured computationally by positing asymmetric learning rates. This asymmetry, motivated by neurobiological and cognitive considerations, has been invoked to explain learning differences across the lifespan as well as a range of psychiatric disorders. Recent theoretical work, motivated by normative considerations, has hypothesized that the learning rate asymmetry should be modulated by the distribution of rewards across the available options. In particular, the learning rate for negative prediction errors should be higher than the learning rate for positive prediction errors when the average reward rate is high, and this relationship should reverse when the reward rate is low. We tested this hypothesis in a series of experiments. Contrary to the theoretical predictions, we found that the asymmetry was largely insensitive to the average reward rate; instead, the dominant pattern was a higher learning rate for negative than for positive prediction errors, possibly reflecting risk aversion.

KEYWORDS: reinforcement learning, multi-armed bandit, decision making

Introduction

Reward prediction error—the discrepancy between observed and predicted reward—plays a central role in many theories of reinforcement learning (Niv and Schoenbaum, 2008; Rescorla and Wagner, 1972; Sutton and Barto, 1990). These theories posit that predictions are incrementally adjusted to reduce the error, with the size of this adjustment determined by a learning rate parameter. Studies have shown that humans differ in the degree to which they learn from positive and negative prediction errors, suggesting asymmetric learning rates (Daw et al., 2002; Frank et al., 2007, 2009; Niv et al., 2012). This asymmetry may arise from the differential response of striatal D1 and D2 dopamine receptors to positive and negative rewards, a hypothesis consistent with individual differences in dopaminergic genes (Frank et al., 2007, 2009) and the effects of dopaminergic medication on learning in patients with Parkinson's disease (Frank et al., 2004; Rutledge et al., 2009) and schizophrenia (Waltz et al., 2007). The learning rate asymmetry also appears to shift across the lifespan: Adolescents learn more from positive prediction errors, while older adults learn more from negative prediction errors (Christakou et al., 2013).

While previous studies have examined differences in the learning rate asymmetry across individuals or medication states, they have generally assumed that the asymmetry is stable over the course of a learning episode. In contrast, Cazé and van der Meer (2013) have recently hypothesized that the asymmetry may dynamically adapt to the distribution of rewards across options. Their hypothesis is based on a normative argument: Asymmetric learning rates can enable an agent to better discriminate reward probabilities, and thereby earn more reward. Importantly, the optimal asymmetry depends on the average reward rate, such that the learning rate for negative prediction errors should be higher than the learning rate for positive prediction errors when the average reward rate is high, and this relationship should reverse when the reward rate is low. Cazé and van der Meer (2013) proposed a meta-

learning algorithm that automatically adapts the asymmetry based on the reward history, and they showed in simulations that this algorithm leads to superior performance compared to an algorithm with fixed learning rates.

The experiments reported in this paper were designed to test the predictions of the adaptive learning rate model. Using a two-armed bandit task, we manipulated the average reward rate across blocks. We then fit several different reinforcement learning models and performed formal model comparison. These models include standard RL models (Daw et al., 2006; Sutton and Barto, 1998), as well as models with asymmetric learning rates (Daw et al., 2002; Frank et al., 2007, 2009; Niv et al., 2012) and variants of the meta-learning model proposed by Cazé and van der Meer (2013). Taken together, these models cover a range of assumptions concerning learning rates that have been proposed in the recent RL literature. Our results show that the learning rate asymmetry is robust across experiments, but this asymmetry does not adapt to the distribution of rewards.

Experiments 1-4

All 4 experiments followed the same procedure, differing only in the reward probabilities (which were not presented explicitly to the participants). On each trial, participants chose one of two options and observed a stochastic binary outcome. The average reward rate was manipulated across blocks, enabling a within-participant comparison of learning rates under different reward rates.

Methods

Participants

A total of 166 participants (ages 23-39) were recruited through the Amazon Mechanical Turk web service: 38 in Experiment 1, 46 in Experiment 2, 45 in Experiment 3, and 37 in Experiment 4. Participants were each paid a flat rate of $0.25. See Crump et al. (2013) for evidence that psychological experiments can be run effectively on Amazon Mechanical Turk.

Procedure

On each trial, participants were shown two colored buttons and told to choose the button that they believed would deliver the most reward. After clicking a button, participants received a binary (0,1) reward with some probability. The probability for each button was fixed throughout a block of 25 trials. There were two types of blocks: low reward rate blocks and high reward rate blocks. On low reward rate blocks, both options delivered reward with probabilities less than 0.5. On high reward rate blocks, both options delivered reward with probabilities greater than 0.5. These probabilities (which were never shown to participants) differed across experiments, as summarized in Table 1. The probabilities were chosen to cover a relatively diverse range and thus enhance the generality of our results.

Each participant played two low reward blocks and two high reward blocks. The button colors for each block were randomly selected, and the assignment of probabilities to buttons was counterbalanced across blocks. Participants were told to treat each set of buttons as independent.

Models

We fit 4 different models to participants' choice data:

1. **Single learning rate**. After choosing option $c_t \in \{1, 2\}$ on trial t and observing reward $r_t \in \{0, 1\}$, the value (reward estimate) of the option is updated according to:

$$V_{t+1}(c_t) = V_t(c_t) + \eta\delta_t, \tag{1}$$

   where $\eta \in [0, 1]$ is the learning rate and $\delta_t = r_t - V_t(c_t)$ is the prediction error. This is the standard temporal difference (TD) model (Daw et al., 2006; Sutton and Barto, 1998) with a single fixed learning rate. For this and subsequent models, all values are initialized to zero.

2. **Dual learning rates**. This model is identical to Model 1, except that it uses two different learning rates, η^+ for positive prediction errors ($\delta_t > 0$) and η^- for negative prediction errors ($\delta_t < 0$). As noted in the Introduction, this model has been proposed by several authors (Daw et al., 2002; Frank et al., 2007, 2009; Niv et al., 2012).

3. **Dual adaptive learning rates**. Like Model 2, this model has separate learning rates for positive and negative prediction errors, but these are adapted automatically by a meta-learning algorithm rather than being treated as fixed parameters. The meta-learning algorithm adapts the learning rates according to:

$$\eta^-_{t+1} = \eta^-_t + \alpha(r_t - \eta^-_t) \tag{2}$$

$$\eta^+_{t+1} = \eta^+_t + \alpha(1 - r_t - \eta^+_t) \tag{3}$$

   These updates are similar to the meta-learning algorithm proposed by Cazé and van der Meer (2013), which estimates the optimal learning rates. Intuitively, these updates will

cause η^- to increase on high reward rate blocks and to decrease on low reward rate blocks, while the opposite pattern will obtain for η^+. The initial values η_1^+ and η_1^- were fit as free parameters.

4. **Extended dual adaptive learning rates**. This model extends Model 3 by allowing the meta-learning rate (α) to vary across positive and negative prediction errors:

$$\eta_{t+1}^- = \eta_t^- + \alpha^-(r_t - \eta_t^-) \tag{4}$$

$$\eta_{t+1}^+ = \eta_t^+ + \alpha^+(1 - r_t - \eta_t^+) \tag{5}$$

where α^- and α^+ are the meta-learning rates for $\delta < 0$ and $\delta > 0$, respectively.[1]

5. **Dual block-specific learning rates**. This model also has separate learning rates for positive and negative prediction errors, but fits them separately for high ($\eta_{\text{high}}^+, \eta_{\text{high}}^-$) and low ($\eta_{\text{low}}^+, \eta_{\text{low}}^-$) reward blocks. Note that participants are not explicitly told what block they are in, so this model is descriptive rather than mechanistic; it is useful insofar as it allows us to test the experimental predictions of Cazé and van der Meer (2013) without making a commitment to a particular meta-learning algorithm. For this reason, we do not include Model 5 in the model comparisons reported below, which are meant to identify a psychologically plausible learning algorithm.

All models use a logistic sigmoid transformation to convert values to choice probabilities:

$$P(c_t = 1) = \frac{1}{1 + e^{-\beta[V_t(1) - V_t(2)]}}, \tag{6}$$

[1] We also fit a version of the dual adaptive learning rate model in which the learning rates are updated according to the Pearce-Hall rule (Pearce and Hall, 1980). However, we found that this model fit the data poorly and for the sake of brevity we will not report these model fits.

where β is a free parameter that governs the exploration-exploitation trade-off. Previous work has shown that this model of choice probability provides a good account of choice variability (Daw et al., 2006).

Model-fitting

Free parameters were estimated for each participant separately using importance sampling (Robert and Casella, 2004). While maximizing likelihood is a more standard parameter estimation technique in the reinforcement learning literature, maximum likelihood has two drawbacks for our purposes. First, it tends to produce parameter estimates with high variance across participants, a consequence of the small amount of data we have per participant. Second, it does not provide an estimate of the marginal likelihood (model evidence), which balances fit against complexity, and is a standard metric for model comparison (see MacKay, 2003, for an overview). While one could use an approximation like the Bayesian Information Criterion (Schwarz, 1978), this approximation is known to over-penalize complexity for small amounts of data. In contrast, importance sampling can produce an arbitrarily accurate estimator of the marginal likelihood, provided we use enough samples.

Letting θ denote the set of parameters, we drew samples $\{\theta_1, \ldots, \theta_M\}$ from a prior distribution $P(\theta)$. We chose $M = 25000$, which yielded stable parameter estimates. Using these samples, the mean of the posterior distribution over parameters is approximated by:

$$\mathbb{E}[\theta|\mathcal{D}] \approx \frac{\sum_{m=1}^{M} P(\mathcal{D}|\theta_m)\theta_m}{\sum_{m=1}^{M} P(\mathcal{D}|\theta_m)}, \tag{7}$$

where $\mathcal{D}$ represents the choice and reward data for a single participant and the likelihood is given by $P(\mathcal{D}|\theta) = \prod_t P(c_t|\theta)$. We assumed that $P(\theta)$ was uniform over the parameter range (for β we restricted this range to $[0.001, 10]$, but our results are not sensitive to this choice).

In order to assess whether participants were choosing non-randomly, we also fit a version of the model that allows β to occupy the range $[-10, 10]$. Although having a negative value of β is non-sensical from a computational point of view (since it induces repulsion from high value choices), this version of the model permits us to test whether β is significantly greater than 0, indicating non-random choice behavior.

To compare models at the group level, we assumed that the marginal likelihood of the data $P(\mathcal{D})$ is a random effect across participants, and submitted these marginal likelihoods to the hierarchical Bayesian method described in Stephan et al. (2009). In brief, this method posits that each participant's data were drawn from one model (among the set of models considered); the probability distribution over models is itself a random variable drawn from a Dirichlet distribution. After estimating the parameters of this Dirichlet distribution, the exceedance probability for each model (the probability that a particular model is more likely than all the other models considered) can be computed and used as a model comparison metric. We used importance sampling to approximate the marginal likelihood for a single participant:

$$\begin{aligned} P(\mathcal{D}) &= \int_{\theta} P(\mathcal{D}|\theta)P(\theta)d\theta \\ &\approx \frac{1}{M}\sum_{m=1}^{M} P(\mathcal{D}|\theta_m). \end{aligned} \tag{8}$$

The marginal likelihood for the group is the product of marginal likelihoods over participants. We computed this group marginal likelihood separately for each model.

Results

The average proportion of correct responses in the last 10 trials of each block was 0.56 across all experiments, significantly greater than chance [$t(165) = 59.63, p < 0.0001$], and significantly greater than the average proportion of correct responses in the first 10 trials of each block [$t(165) = 2.48, p < 0.05$]. This low level of performance reflects the difficulty of the task, which only gives participants 25 trials to distinguish probabilities that are separated by 0.2 (Experiments 1 and 2) or 0.1 (Experiments 3 and 4). To confirm that participants were treating the blocks as independent, we correlated the performance metric measured on neighboring blocks. After Fisher z-transforming these correlations, we found that they were not significantly greater than 0 across all experiments ($p = 0.61$).

Turning to model-based analyses of the data, we sought to confirm that the class of models described above was sufficiently rich to capture choice probabilities in our experiments. Figure 1 shows empirical and predicted choice probabilities for each model as a function of the value difference, $V(1) - V(2)$. As these results demonstrate, all the models do a good job capturing the choice probability curve (we excluded Model 5 from this comparison, since it is not a mechanistic model of the task, but the results look similar). We next asked whether participants effectively exploited their learned knowledge about the probabilities (i.e., choosing non-randomly), by fitting a version of the models that allows β to be less than 0 (see Methods). We found that β was significantly greater than 0 [$t(165) = 12.79, p < 0.0001$]. Thus, participants appear to be choosing non-randomly. All the following analyses use the model variants which restrict β to the range $[0.001, 10]$.

We then addressed the central question of the paper: do learning rates adapt to the distribution of reward? The parameter estimates for Model 5 and exceedance probabilities for Models 1-4 are shown in Figure 2 (mean parameter estimates for all models are displayed in Table 2). Across all 4 experiments, a fairly consistent picture emerges from

the Model 5 parameter estimates: The learning rate for negative prediction errors (η^-) is greater than the learning rate for positive prediction errors (η^+). We confirmed this observation statistically by running an ANOVA with reward rate (high vs. low), prediction error type (positive vs. negative), and experiment as factors (note that reward rate and prediction error type here refer to descriptors of the learning rate parameters). We found an effect of prediction error type [$F(1, 162) = 39.02, p < 0.0001$], and an effect of reward rate [$F(1, 162) = 5.02, p < 0.05$]. The effect of reward rate was primarily driven by the results of Experiment 1; when examined individually, only Experiment 1 showed a significant effect of reward rate [$F(1, 37) = 5.73, p < 0.05$]. Importantly, we found no interaction between prediction error type and reward rate ($p = 0.12$), disconfirming the predictions of Cazé and van der Meer (2013). We also found no effect of experiment ($p = 0.94$), indicating that small variations in the reward probabilities do not exert a significant effect on the learning rate asymmetry.

Our formal model comparison, using the method described in Stephan et al. (2009), showed generally strong support for a model with fixed separate learning rates for positive and negative prediction errors (Model 2). The only exception was Experiment 3, where the exceedance probability for Model 2 was relatively low. This appears to be a consequence of the fact that no learning rate asymmetry was found for the High reward condition, as shown by an analysis of the learning rates for Model 5 ($p = 0.53$). In this case, the lack of a reliable learning rate asymmetry in Experiment 3 favored the simpler Model 1 (which has one less free parameter). Nonetheless, when the marginal likelihoods for all experiments were pooled together, the exceedance probability for Model 2 was indistinguishable from 1. In no case did we find appreciable support for Models 3 or 4, meta-learning models similar to the one suggested by Cazé and van der Meer (2013).

One issue in interpreting these results is that the meta-learning models are more complex

(i.e., have more parameters) than the other models, and hence they will be more strongly penalized by the model comparison metric. This possibility is suggested by Figure 1, where Models 3 and 4 appear to have a better fit to the choice probability data. To address this issue, we fit a version of the meta-learning models in which the learning rates are initialized to 0 and updated before the value update, so that the initial value is proportional to the first reward (in the case of the negative learning rate), or proportional to 1 minus the first reward (in the case of the positive learning rate). This eliminates two free parameters from the models. Our model comparison results were largely the same as shown in Figure 2, indicating that the lower model evidence for the meta-learning models is not simply due to a complexity penalty.

It is possible that some participants were poorly fit by Model 5, which could explain the absence of a learning rate asymmetry. To address this possibility, we correlated the evidence for Model 5 with the interaction effect computed by the ANOVA. For all 4 experiments, we failed to find a significant correlation ($p > 0.49$), indicating that participants who are better explained by the model do not show a stronger learning rate asymmetry.

Another potential concern is that the experiments are insufficiently powered to discover a learning rate asymmetry should one exist. To address this concern, we performed a simulation study. For each experiment and each model, we generated simulated data from artificial agents with parameters drawn from a normal distribution fitted to the empirical parameter estimates.[2] The data set was the same size as the actual experiments (4 blocks, 25 trials per block), with the same number of participants. We then fit each model to the simulated data and examined the exceedance probabilities. Figure 3 (top) shows that the exceedance probability for the correct model (i.e., the one that generated the data) was very close to 1 across all experiments. Thus, our experimental design and model-fitting procedure can re-

[2]Largely the same results were obtained with parameters drawn randomly from the prior (i.e., uniformly within the parameter bounds).

cover the correct model with very high accuracy. We also examined the accuracy with which parameters can be recovered. As shown in Figure 3 (bottom), the correlation between the inferred and ground truth parameters always exceeded 0.84, and the median correlation was 0.95, demonstrating that subtle variations in parameter values can be recovered accurately. We conclude that the experiments are indeed sufficiently powered to discover a learning rate asymmetry should one exist.

Discussion

The results of four experiments provide evidence for reinforcement learning models with separate learning rates for positive and negative prediction errors (Christakou et al., 2013; Frank et al., 2004, 2007, 2009; Niv et al., 2012; Waltz et al., 2007). In particular, the negative learning rate was generally higher than the positive learning rate, consistent with the results of Niv et al. (2012). This may reflect risk aversion: a higher negative learning rate drives choices away from risky options (Mihatsch and Neuneier, 2002).

The results failed to support a recent normative model proposed by Cazé and van der Meer (2013), according to which the learning rate asymmetry should adapt to the distribution of rewards. Instead, we found that the learning rate asymmetry is mostly stable over a variety of different reward distributions. Because we have only studied choices between two options with binary gains, more research will be required to evaluate the generality of our conclusions.

Beyond learning rate asymmetries, recent research on reinforcement learning has lead to a plethora of other ideas about learning rates, include dynamic volatility-sensitive adjustment (Behrens et al., 2007), selective attention (Dayan et al., 2000), multiple timescales (Bromberg-Martin et al., 2010), and neuromodulatory control (Doya, 2002). Some of these

ideas have deep roots in associative learning theory (e.g., Mackintosh, 1975; Pearce and Hall, 1980). Theorists are now faced with the challenge of formalizing how these disparate ideas fit together. Toward this end, it is crucial to ascertain which theoretical predictions are robust across experimental manipulations. The contribution of the present study is to sharpen our empirical understanding of the factors governing learning rates, and to show how this can aid in whittling down the complex tangle of assumptions underpinning contemporary reinforcement learning theory.

Acknowledgments

This work was supported by a postdoctoral fellowship from the MIT Intelligence Initiative. I am grateful to Yael Niv for helpful discussions.

References

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10:1214–1221.

Bromberg-Martin, E. S., Matsumoto, M., Nakahara, H., and Hikosaka, O. (2010). Multiple timescales of memory in lateral habenula and dopamine neurons. *Neuron*, 67:499–510.

Cazé, R. D. and van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, 107:711–719.

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., and Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, 25:1807–1823.

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, 8:e57410.

Daw, N. D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, 15:603–616.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441:876–879.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3:1218–1223.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, 15:495–506.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, 12:1062–1068.

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

Frank, M. J., Seeberger, L. C., and O'Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in Parkinsonism. *Science*, 306:1940–1943.

MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82:276–298.

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290.

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

Niv, Y. and Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, 12:265–272.

Pearce, J. M. and Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87:532–552.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. and Prokasy, W., editors, *Classical Conditioning II: Current Research and theory*, pages 64–99. Appleton-Century-Crofts, New York, NY.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., and Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in parkinson's patients in a dynamic foraging task. *The Journal of Neuroscience*, 29:15104–15114.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46:1004–1017.

Sutton, R. and Barto, A. (1990). Time-derivative models of pavlovian reinforcement. In Gabriel, M. and Moore, J., editors, *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pages 497–537. MIT Press.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Waltz, J. A., Frank, M. J., Robinson, B. M., and Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, 62:756–764.

Tables

Table 1: Design of experiments. The numbers in columns 2 and 3 represent the reward probabilities for each action in a block.

Experiment	**Low reward blocks**	**High reward blocks**
1	0.2, 0.4	0.6, 0.8
2	0.1, 0.3	0.7, 0.9
3	0.1, 0.2	0.8, 0.9
4	0.2, 0.3	0.7, 0.8

Table 2: Parameter estimates (mean across participants) for all models.

Experiment	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
1	$\beta = 3.24$ $\eta = 0.47$	$\beta = 4.03$ $\eta^+ = 0.37$ $\eta^- = 0.57$	$\beta = 2.92$ $\eta_1^+ = 0.41$ $\eta_1^- = 0.49$ $\alpha = 0.3$	$\beta = 2.95$ $\eta_1^+ = 0.40$ $\eta_1^- = 0.48$ $\alpha^+ = 0.28$ $\alpha^- = 0.42$	$\beta = 3.77$ $\eta_{\text{low}}^+ = 0.39$ $\eta_{\text{low}}^- = 0.50$ $\eta_{\text{high}}^+ = 0.45$ $\eta_{\text{high}}^- = 0.58$
2	$\beta = 2.82$ $\eta = 0.40$	$\beta = 3.08$ $\eta^+ = 0.37$ $\eta^- = 0.50$	$\beta = 2.08$ $\eta_1^+ = 0.43$ $\eta_1^- = 0.48$ $\alpha = 0.37$	$\beta = 1.91$ $\eta_1^+ = 0.44$ $\eta_1^- = 0.49$ $\alpha^+ = 0.39$ $\alpha^- = 0.44$	$\beta = 2.61$ $\eta_{\text{low}}^+ = 0.42$ $\eta_{\text{low}}^- = 0.49$ $\eta_{\text{high}}^+ = 0.43$ $\eta_{\text{high}}^- = 0.5$
3	$\beta = 4.32$ $\eta = 0.43$	$\beta = 4.41$ $\eta^+ = 0.42$ $\eta^- = 0.47$	$\beta = 3.45$ $\eta_1^+ = 0.42$ $\eta_1^- = 0.47$ $\alpha = 0.31$	$\beta = 3.46$ $\eta_1^+ = 0.41$ $\eta_1^- = 0.49$ $\alpha^+ = 0.34$ $\alpha^- = 0.42$	$\beta = 4.27$ $\eta_{\text{low}}^+ = 0.42$ $\eta_{\text{low}}^- = 0.50$ $\eta_{\text{high}}^+ = 0.47$ $\eta_{\text{high}}^- = 0.49$
4	$\beta = 3.04$ $\eta = 0.43$	$\beta = 3.66$ $\eta^+ = 0.35$ $\eta^- = 0.55$	$\beta = 2.49$ $\eta_1^+ = 0.41$ $\eta_1^- = 0.50$ $\alpha = 0.34$	$\beta = 2.49$ $\eta_1^+ = 0.40$ $\eta_1^- = 0.49$ $\alpha^+ = 0.34$ $\alpha^- = 0.47$	$\beta = 3.18$ $\eta_{\text{low}}^+ = 0.37$ $\eta_{\text{low}}^- = 0.56$ $\eta_{\text{high}}^+ = 0.45$ $\eta_{\text{high}}^- = 0.52$

Figure captions

Figure 1: Choice probabilities. Each panel shows the average human and model probabilities of choosing option 1, plotted as a function of the value difference, $V(1) - V(2)$. On each trial, we recorded whether or not a participant chose option 1, along with the estimated value difference on that trial for each model; the plotted choice probabilities represent averages across trials. Data are combined across all 4 experiments. Note that the data are the same in all 4 panels, but the curves appear slightly different because they are binned based on the model-based values (which differ across panels). Also note that value differences can exceed the differences in reward probabilities because the values are updated incrementally and hence can cover the entire $[0, 1]$ interval.

Figure 2: Model-based analyses. (*Top*) Posterior mean parameter estimates for Model 5 (dual block-specific learning rate model). Error-bars represent within-subject standard errors of the mean. (*Bottom*) Exceedance probabilities for Models 1-4.

Figure 3: Simulation study. For each experiment, simulated data generated by one model were fit by all the models. (*Top*) Exceedance probabilities for each model combination. The rows correspond to the ground truth model, and the columns correspond to the model used to fit the data. White indicates an exceedance probability of 0; black indicates an exceedance probability of 1. In all cases, the exceedance probability of the correct (data-generating) model was indistinguishable from 1. (*Bottom*) Correlation between the ground truth and inferred parameters for each model.

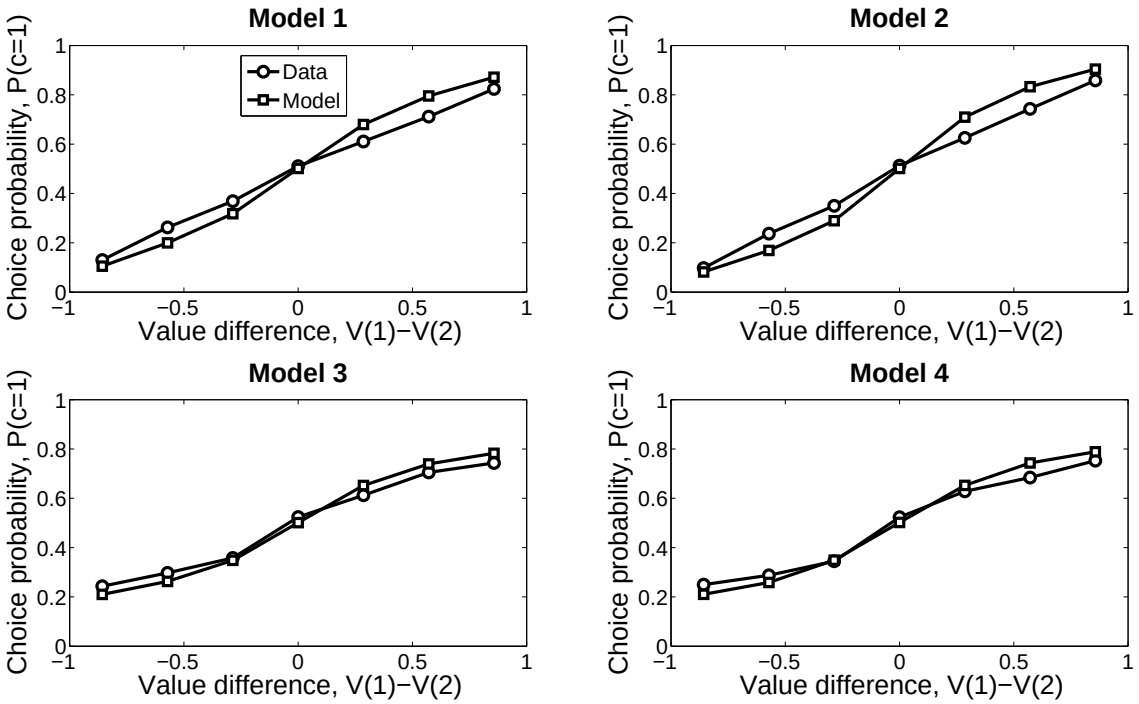

Figure 1: Choice probabilities.

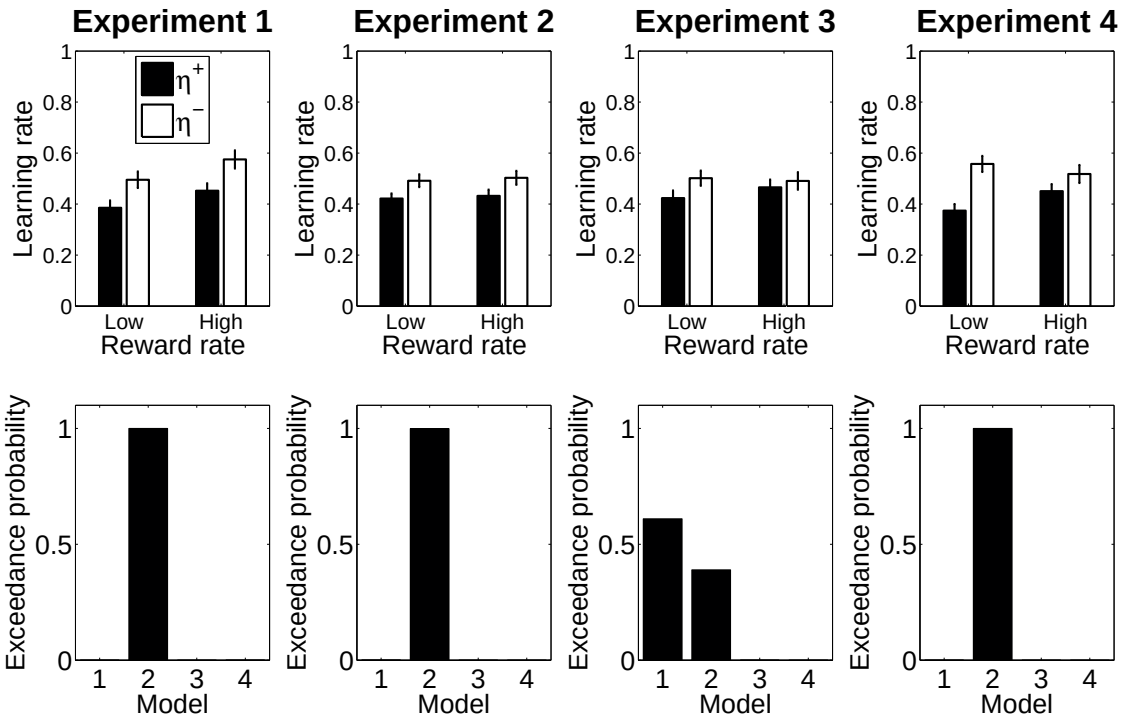

Figure 2: Model-based analyses.

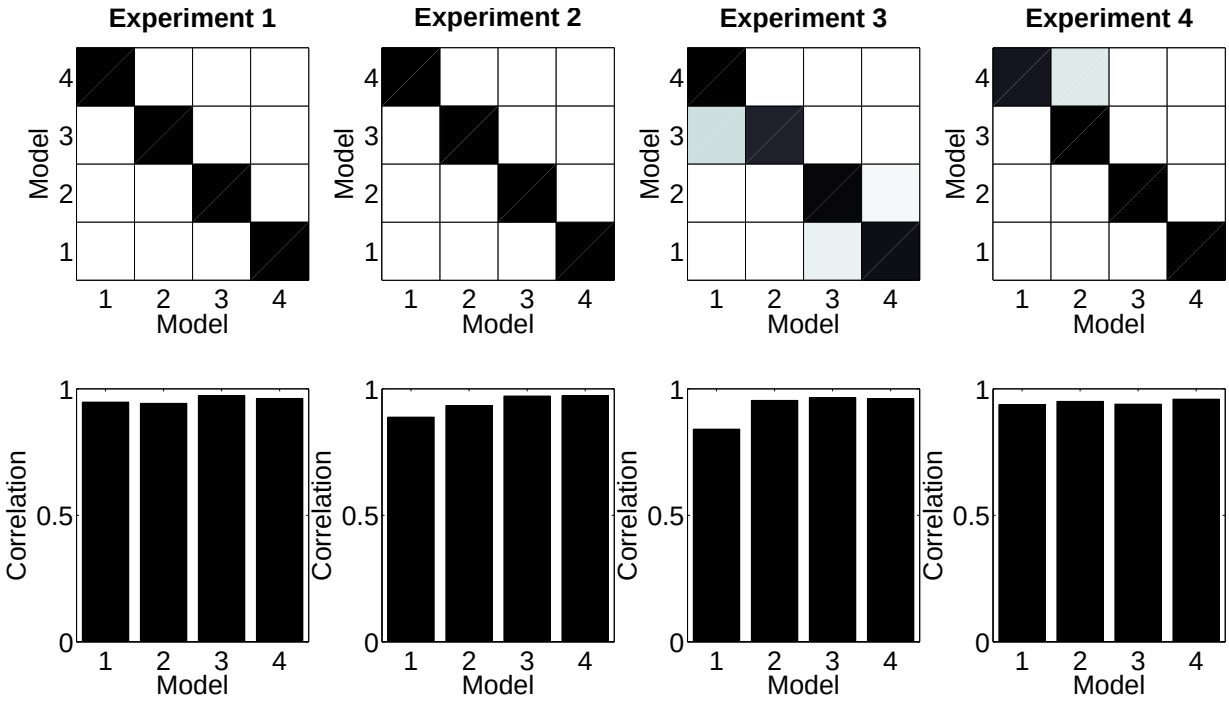

Figure 3: Simulation study.