A Case of Divergent Predictions Made by Delta and Decay Rule Learning Models

Darrell A. Worthy (worthyda@tamu.edu)
Department of Psychological & Brain Sciences, 4235 TAMU
College Station, TX 77843-4235 USA

A. Ross Otto (ross.otto@mcgill.ca)
Department of Psychology, 2001 McGill College Ave.
Montreal, QC H3A 1G1 Canada

Astin C. Cornwall (acornwall@tamu.edu)
Department of Psychological & Brain Sciences, 4235 TAMU
College Station, TX 77843-4235 USA

Hilary J. Don (hdon7006@uni.sydney.edu.au)
School of Psychology, Griffith Taylor Building (A19), University of Sydney
NSW 2006, Australia

Tyler Davis (tyler.h.davis@ttu.edu)
Department of Psychological Sciences, MS 2051 Psychology Building
Lubbock, TX 79409-2051 USA

Abstract

The Delta and Decay rules are two learning rules used to update expected values in reinforcement learning (RL) models. The delta rule learns *average* rewards, whereas the decay rule learns *cumulative* rewards for each option. Participants learned to select between pairs of options that had reward probabilities of .65 (option A) versus .35 (option B) or .75 (option C) versus .25 (option D) on separate trials in a binary-outcome choice task. Crucially, during training there were twice as AB trials as CD trials, therefore participants experienced more cumulative reward from option A even though option C had a higher average reward rate (.75 versus .65). Participants then decided between novel combinations of options (e.g, A versus C). The Decay model predicted more A choices, but the Delta model predicted more C choices, because those respective options had higher cumulative versus average reward values. Results were more in line with the Decay model's predictions. This suggests that people may retrieve memories of cumulative reward to compute expected value instead of learning average rewards for each option.

Keywords: reinforcement learning, delta rule, decay rule, prediction error, base rates, probability learning

Introduction

The Delta Rule model is a simple learning model that has become the default model of behavior in simple choice tasks where participants learn via feedback. Delta-based learning models have been applied to a variety of learning contexts including reward/value learning, associative conditioning, and category learning (e.g. Sutton & Barto, 1981; 1998; Williams, 1992; Jacobs, 1988; Gluck & Bower, 1988; Rumelhart & McClelland, 1986; Widrow & Hoff, 1960; Rescorla & Wagner, 1972; Busemeyer & Stout, 2002; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

In the present study, we examine RL model predictions for two alternative choice tasks in which participants receive binary rewards based on fixed probabilities tied to each option. Take for example a hypothetical task in which participants learn to choose between an option A that is rewarded 65% of the time and option B that is rewarded 35% of the time. Delta rule models can accurately learn which options are more valuable in this scenario by tracking the recency-weighted average reward participants receive. If rewards (r) are coded as 1 when a reward is given and 0 when a reward is not given then the expected value (EV) for each *j* option is computed by the Delta rule on each *t* trial as:

$$EV_j(t+1) = EV_j(t) + \alpha \cdot (r(t) - EV_j(t)) \cdot I_j \qquad (1)$$

Where I_j is simply an indicator value that is set to 1 if option *j* is selected on trial *t,* and 0 otherwise. Critically, the update function on the delta rule means that expected values are only updated for the chosen selection. If participants choose A for an AB pair, they update their information about A, but not B. The portion of Equation 1 in parentheses is known as the prediction error, and it is modulated by the learning rate parameter ($0 \leq \alpha < 1$). Higher values of α indicate greater weight to recent outcomes, while lower values indicate less weight to recent outcomes. When α=0 no learning takes place and expected values remain at their starting points, and when α=1 expected values are equal to the last outcome received for each option.

The predicted probability that option *j* will be chosen on trial *t*, $P\left|C_j(t)\right|$, is calculated using a Softmax rule:

$$P\left|C_j(t)\right| = \frac{e^{\beta \cdot EV_j(t)}}{\sum_1^{N(j)} e^{\beta \cdot EV_j(t)}} \qquad (2)$$

Where $\beta = 3^c - 1\ (0 \leq c)$, and c is an inverse temperature parameter that determines how consistently the option with the higher expected value is selected (Yechiam & Ert, 2007). When c=0 choices are random, and as c increases the option with the highest expected value is selected most often.

The Delta rule model is sometimes called the Basic RL model (Collins & Frank, 2012; Worthy, Maddox, & Markman, 2007). In their 1981 paper Sutton and Barto referred to it as the Rescorla-Wagner/Widrow-Hoff rule after noting that the learning rules presented in those two papers were identical, and also noting how common similar Delta-based learning mechanisms were to a wide variety of models. A canonical finding in the neuroscience literature is that prediction errors are correlated with activation in the ventral striatum and medial prefrontal cortex (Schultz, Dayan, & Montague, 1997, Schultz & Dickinson, 2000; McClure, Berns, & Montague, 2002; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Samanez-Larkin, Worthy, Mata, McClure, & Knutson, 2014).

In the AB choice scenario described above, the Delta Rule model will learn to select the option with the higher expected value as long as both of its parameters are non-zero. For the purpose of this paper, the key component of the Delta rule model is that it learns recency-weighted *average* rewards provided by each option. In a binary outcome task with rewards coded as 0 or 1 this will roughly equate to the average probability of reward receipt provided by each option.

We contrast the Delta Rule model with a separate model for explaining the updating of expected values, the Decay Rule model. The Decay Rule model was developed by Erev and Roth (1998), and further examined by Yechiam and colleagues (Yechiam & Busemeyer, 2005; Yechiam & Ert, 2007). So far, it has enjoyed less wide spread adoption, but is a core component of the Prospect Valence Learning model (PVL; Ahn et al., 2008) model of the Iowa Gambling Task (although Delta-based versions of PVL exist as well; Worthy, Pang, & Byrne 2013). Outside of the Iowa Gambling Task literature, however, most modeling efforts focus more on utilizing the delta rule rather than the decay rule.

The Decay Rule model also tracks expected values, but it does so without utilizing prediction errors. Specifically, on each *t* trial the EV for each *j* option is updated according to:

$$EV_j(t+1) = EV_j(t) \cdot A + r(t) \cdot I_j \qquad (3)$$

As in Equation 1, I_j is an indicator variable that is set to 1 if option *j* was selected on trial *t*, and 0 otherwise. A $(0 \leq A \leq 1)$ is a decay parameter, and $r(t)$ is the reward given on each trial. While Equations 1 and 3 share some similarities the key difference is that the Decay rule tracks the recency-weighted *cumulative* reward provided by each option, whereas the Delta rule tracks the recency-weighted *average* reward provided by each option. In the example AB choice task described above with reward probabilities of .65 and .35 the Decay rule's EVs will not converge to the average reward provided by each option, but will instead increment to larger values as one option is selected and rewarded more often than the other. This incremental process is balanced by the decay parameter which causes all options to decay in value on each trial, particularly options that are not selected. The Decay rule will also predict greater perseveration, or repeated sampling of options chosen on previous trials, provided that rewards are given instead of losses, because selected options will increase in value, while non-selected options will decrease in value (Worthy et al., 2013).

It's also worth noting that our view of the Decay model as learning a combined memory of previous rewards is similar to the theory of Decision by Sampling where value is derived from sampling previously experienced items in memory (Stewart, Chater, & Brown, 2006). It's also similar to the idea behind the Instance-based learning model (IBL; Gonzalez & Dutt, 2011). While the Decay model does not store specific instances, its expected values should be very similar to the expected value that come fromblending instances in the IBL. While it is beyond out present scope, future work should compare these models more directly.

To summarize, for the present study, the key difference between the Delta and Decay rule models is the Delta Rule model learns average rewards provided by each option, while the Decay Rule model learns cumulative rewards provided by each option, both weighted by recent action selection history. The Decay rule model will also predict more perseveration than the Delta Rule model because it decays non-chosen actions.

One question that emerges is which model is better at accounting for human behavior in simple decision-making tasks like the hypothetical AB task described above? Previous work suggests that the Decay Rule model usually provides better fits to data than the Delta Rule model because it accounts for perseveration better than the Delta Rule model; however, the Delta Rule model has shown superior generalization to other tasks precisely because it does not give as much weight to past choices, or perseverative tendencies, and gives more weight to past outcomes compared to the Decay Rule model (Steingroever, Wetzels, & Wagenmakers, 2014). Here, we provide novel evidence to differentiate between the two models by pitting them against each other in an experiment for which they make qualitatively different predictions about choice behavior.

Manipulating Base Rates to Compare the Two Models

The present experiment is based on early research on probability learning. One critical question from the probability learning literature is whether people learn probabilities per se when performing simple binary outcome tasks like the one describe above, or whether they simply store memories for past rewards which are later translated into probability judgments (Estes, 1976). The key insight in this study was that memory and probability learning could be disentangled by manipulating choice base rates; models that assume people learn to track probabilities per se will be unaffected by base rate manipulations, whereas memory based models will form stronger memory for higher frequency options. Results from three experiments supported the idea that participants use memories for outcomes instead of tracking probabilities directly (Estes, 1976).

Although not anticipated at the time, this question maps onto key differences between the Delta Rule and Decay Rule models. The Delta Rule model learns the probability of receiving a reward for each option, and the Decay Rule model tracks how often each option has provided a reward. The Decay Rule model's EVs can be thought of as a summation of how often participants have been rewarded after selecting each option, which decays in memory.

To investigate this difference in how the two models may be affected by base rate differences in choice availability, we designed a task resembling the AB choice scenario used to introduce the two models above. The task involves four choices A-D, that are learned in pairs; participants are shown A versus B, make a choice, and receive feedback, or they are shown C versus D, and they make a choice and receive feedback. These trials are interspersed and the probabilities of reward receipt associated with each option are: [.65, .35, .75, .25] for options A-D respectively. Thus A is the optimal choice on AB trials, and C is the optimal choice on CD trials, and the optimal choice overall.

The key base rate manipulation is, over 150 training trials, there are 100 AB trials, and only 50 CD trials. Thus AB trials occur twice as often as CD trials. This manipulation should not affect the Delta Rule's EVs because this model learns average rewards. The EVs for the Delta Rule model should roughly correspond to the actual reward probabilities provided by each option. However, for the Decay Rule model the expected values should be affected by the differences in base rates. Because the Decay Rule model learns cumulative rewards provided by each option, or a decaying memory of total reward, the EV for option A should be the highest, most notably higher than the EV for option C, which has a higher probability of reward receipt (.75 versus .65).

To test these qualitatively different predictions between the Delta and Decay models, we use a post-learning test phase in which participants choose amongst novel pairs of choices that they were not previously trained on (CA, CB, AD, BD). CA trials are of most interest, as the Delta Rule model should predict more C choices and the Decay Model should predict more A choices.

To verify these predictions, we simulated this task with the Delta and Decay Rule models, and examined their predictions for the EVs of each option (A, B, C, and D) at the end of training. We simulated 1000 data sets for each combination of α or A and c. α or A varied from 0 to 1 in increments of .05, and c varied from 0 to 5 in increments of .25. We then examined the average expected values for each option for the Delta and Decay Rule models across the 1,000 simulations for each parameter combination.

Figure 1 plots the average probability of selecting option C on CA trials for each parameter combination for the Delta (a) and Decay (b) models. The key result can be seen by looking at the scales on the right-hand side of each figure. Delta Model predictions range from .5 to about .75, while Decay model predictions range from about .15 to .55. Overall it seems clear that the Delta model generally predicts more C choices, while the Decay model generally predicts more A choices.

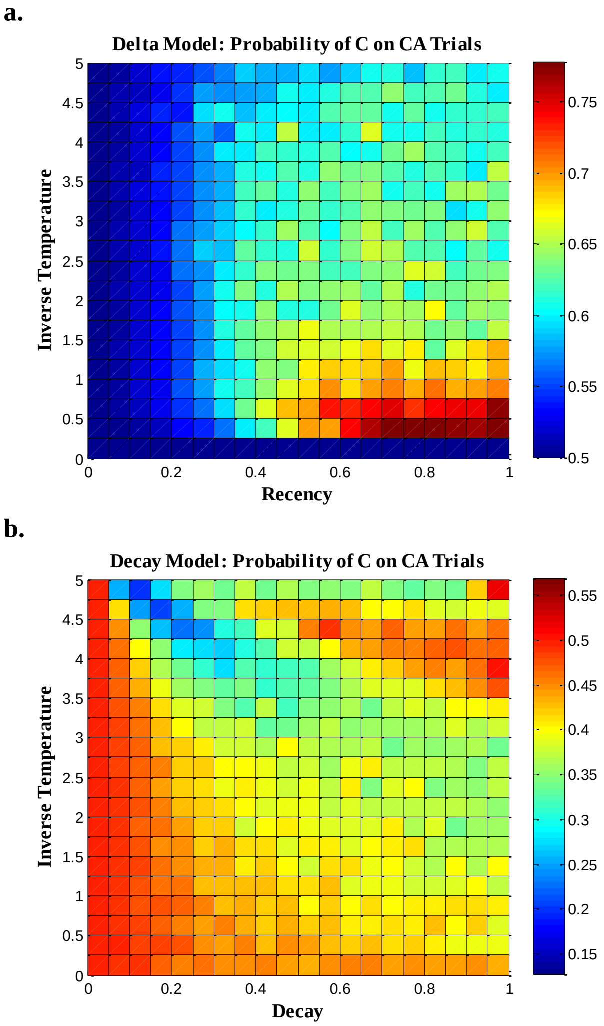

Figure1 (a): Expected values at the end of training for each model, averaged across all parameter combinations. (b): Probability of selecting the optimal option predicted by each model, averaged across all parameter combinations.

The predictions of the Delta Rule model are more optimal because option C has an objectively higher value than option A. However, the question is what will human subjects prefer? Given the close alignment between our present question about expected value and Estes (1976) work on probability learning, we predicted that participants would not learn EVs directly and instead would make choices based on memories of rewarding events, as predicted by the Decay Rule model.

Experiment

Method

Participants

Thirty-three participants from Texas A&M University participated in the experiment for partial fulfillment of a course requirement. The Internal Review Board approved the experiment, and participants gave informed consent.

Materials and Procedure

Participants performed the experiment on PCs using Matlab software with Psychtoolbox version 2.54. The experiment was identical to the simulation described above. Of note, participants did receive feedback during the test phase. Future work should replicate the experiment with no feedback during test.

Results

We computed the proportion of optimal choices made for each trial type. These are shown in Figure 2. The first letter of each pair is the optimal choice (e.g. C for CA trials). We conducted one sample t-tests for each pair using the ratio of their objective reward probabilities. For example, for CA trials we used a test probability of .75/(.75+.65)=.5357. We compared against this baseline because it allowed us to test for a bias that differed from the bias from one option being objectively more rewarding than the other. As seen in Figure 2 participants selected option C on CA trials only 40% of the time. This is significantly different from the test value of .5357, *t(*32)=-3.16, *p=*.003, *BF*=10.86. The median of the distribution was .40, with .24 and .56 as the 25th and 75th percentiles. 15% of participants selected option C more than 70% of the time compared to 30% who selected option A more than 70% of the time. These results are more consistent with the Decay model than the Delta model, although a small group of participants appeared to have learned that C had a higher probably of reward.

For CB trials participants selected option C only 42.6% of the time. This is far from 68.8% of trials which corresponds to the ratio of the two options' reward probabilities, *t(*32)=-5.31, *p*<.001, *BF*=2,540; this BF corresponds to *extreme* evidence for the alternative hypothesis. For AD trials participants selected option A on 77.9% of trials, which is slightly more than the .722 ratio of those options' reward probabilities, *t(*32)=2.20, *p*=.041, *BF*=1.35. Similarly, for BD trials, option B was selected (67%) slightly more often than the default value of 58.3% of trials predicted by the ratio of the options' reward probabilities, *t(*32)=2.29, *p*=.029, *BF*=1.79.

The left side of Figure 2 shows the proportion of optimal choices made during training. On AB trials option A was selected 72.2% of the time which is more than the 65% of trials predicted by the ratio of the options' reward probabilities, *t(*32)=3.25, *p*=.003, *BF*=13.30. On CD trials option C was selected (70%) slightly less often than the 75% from the options' reward probability ratio, *t(*32)=-2.17, *p*=.038, *BF*=1.46. These clearly demonstrate that participants learned which option was more rewarding, but participants *maximized* more on AB trials than on CD trials by picking the optimal choice more than probability matching would dictate.

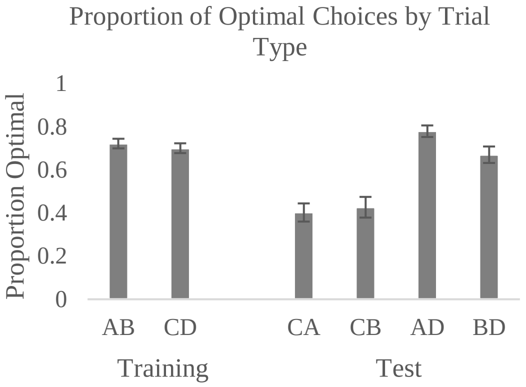

Figure 2: Proportion of objectively optimal choices for each trial type. The first letter listed in each pair represents the optimal choice.

We also examined the observed reward probabilities and total average reward. For options A-D the proportion of trials that participants received reward after selecting each option was [.646; .389; .735; .217]. The total rewards for options A-D, averaged across subjects, were [68.9; 23.1; 41.2; 6.2]. This demonstrates that option A was rewarded on more trials than option C, even though option C was rewarded more on average, each time it was selected.

Theoretical Analysis

We next fit the Delta Rule and Decay Rule models to participants' data individually by maximizing the log-likelihood of the model's next-step-ahead predictions on each trial. EVs were initialized at .5 and updated according to Equations 1 or 3 above, for the Delta and Decay Rule models, respectively. Equation 2 was used for both models to compute action selection probabilities. Fits were obtained using Matlab's fminsearch algorithm with 100 random starting points per subject. The average BIC for the Delta Rule model 308, while the average for the Decay Rule model was 275 (lower is better). This difference in BIC suggests extreme evidence that the Decay model better accounts for the data than the Delta model, with a *BF* of over 22 million (Wagenmakers, 2007). Data from 28 of 33 participants (85%) was best fit by the Decay Rule model, which is significantly different than 50% by a binomial test, *p*<.001. However, McFadden's pseudo R^2 computed against a completely random null model (*ln*(.50) *249) was only .22 for the Decay model and .12 for the Delta model. Thus, there is still a great deal of variance in behavior that is not accounted for by either model.

For the Delta rule model the average learning rate parameter (α) was .30 (*SD*=.30), and the average inverse temperature parameter (*c*) was 1.72 (*SD*=1.61). For the Decay Rule model the average decay parameter (*A)* was .85 (*SD*=.25), and the average inverse temperature parameter value was .37 (*SD*=.27). For the Delta Rule model the average best fitting expected values for options A-D averaged across all trials and then across participants were [.60; .31; .66; .30]. The same values for the Decay Rule model were [9.50; 4.58; 4.67; 1.09]. This is important because even though parameters were allowed to freely vary for each participant, the Delta Rule model had the highest expected value for option C, while the Decay Rule model had the highest expected value for option A. This result reinforces our assertion that the two models make opposite predictions for test trials that are relatively consistent across the parameter spaces of the models. Despite the differences in base rates the Delta rule model cannot predict a higher expected value for option A than option C, which is likely one reason why it cannot provide as good of fit to the data.

Discussion

Here we have demonstrated that the Delta and Decay rule models make divergent predictions about learning options' values when their base rates differ. The Delta rule model learns a recency-weighted average reward associated with each option, while the Decay rule model learns the recency-weighted cumulative reward provided by each option. In our simulations, we showed that the Delta rule model prefers options based on the learned probability of reward assigned to each option, while the Decay rule model prefers options that have provided more reward overall on past trials. In our experiment, the critical test between the two models was whether human participants preferred option A, the more frequently rewarded option, or C, the option with the highest reward probability. The Decay model predicted more A choices because participants had received more reward overall from option A due to its higher base rate. The Delta model favored option C because it had a higher average rate of reward, even though it was available as a choice alternative less often. Most participants selected option A more than option C on these critical trials, in support of the Decay model's predictions. This suggests that they based their decisions more on how often each item was associated with reward in memory, rather than on a learned estimate of the probability of reward receipt.

Delta-based learning is commonly used to model the learning of action values from experience in diverse fields such as Psychology (Otto &Love, 2010), Computer Science and Neuroscience (e.g., McClure et al., 2003). Given the predominance and prevalence of this formalism—and the assumptions it makes about how value learning unfolds—it is important to validate that the Delta learning model does indeed provide the best account of learning, as operationalized by choice behavior or with neural activity. Here we provide a clear case where the Decay rule appears to provide a better account of human behavior than the Delta rule. Participants' choices were more in line with how often they had been rewarded for each option in total rather than on average. This is in line with theories that suggest that people do not learn probabilities of reward directly, but they store instances of reward associated with each option in memory and then translate these into choice probabilities that guide their behavior (Gonzalez & Dutt, 2011; Stewart et al., 2006).

Although, our results support the Decay model there is still an extensive body of work that supports predictions made by the Delta rule model (Rangel et al., 2008). A major finding is that prediction errors from the Delta model are correlated activation of the ventral striatum (e.g. Hare et al., 2008; McClure et al., 2003). One future line of work we are currently pursuing is to examine ways in which a Decay rule model might generate a prediction error. We believe there are possible candidate Decay rule model prediction errors, but future work is needed to examine how these metrics would compare to prediction errors from Delta rule models. Additional work can also be undertaken to identify whether neural activation in RL tasks, as measured by fMRI, is better characterized by Decay rule versus Delta rule prediction errors and expected values. This could potentially be addressed with extant data sets, applying model-based fMRI using each model. The Delta and Decay rule models make similar predictions in a number of situations. Therefore, it is possible that some of the key findings that have been supported predictions made by Delta rule models over the past several decades could also be predicted by Decay models. Alternatively, there may be situations where Decay models make predictions that do not align with human behavior or cognition (e.g. Steingroever et al., 2014). Finally, Bayesian versions of the Delta rule model have recently been developed to account for uncertainty in addition to expected value (Gershman, 2015). A Bayesian Decay model may account for more variance in behavior, such as the exploration/exploitation tradeoff, than the simple variant we used here and should be explored in future work.

It is worth noting that modifications could be made to the Delta rule model by allowing EVs to decay on each trial or by adding a perseveration component (Worthy & Maddox, 2014). Thus, a modified Delta rule model with additional parameters may be able to account for the pattern of behavior we observed, although such a model would still not allow rewards to cause a cumulative increase in EV. Here, we have focused on parsimonious models that represent default models of each type in order to generate specific predictions regarding whether people learn average reward probabilities or memories for rewarding events, but developing a more complex model with additional parameters may be useful in other cases. A major point of Estes' 1976 paper is that "probability learning is in a sense a misnomer," or that people do not directly learn reward probabilities. The Delta Rule model tacitly assumes probability learning, which is inconsistent with the data from most of our participants. It will be necessary to replicate and extend this work, and further test the key predictions made about learning and behavior by different formal models.

Acknowledgements

This work was supported by NIA grant AG043425 to DAW. We thank all of the research assistants in the Worthylab who assisted in data collection.

References

Ahn, W.Y., Busemeyer, J.R., Wagenmakers, E.J., & Stout, J.C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32,* 1376-1402.

Busemeyer, J.R., & Stout, J.C. (2002). A contribution of cognitive decision model to clinical assessment: Decomposing performance on the Bechara Gambling Task. *Psychological Assessment, 14,* 253-262.

Collins, A.G.E., & Frank, M.J. (2012). How much of reinforcement learning is working memory not reinforcement learning? A behavioral, computational and neurogenetic analysis. *European Journal of Neuroscience, 35,* 1024-1035.

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

Erev, I., & Roth, A.E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88,* 848-881.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37-64.

Gershman, S.J. (2015). A unifying probabilistic view of associative learning. *PLoS ComputationalBiology, 11,* e1004567

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 128,* 309-331.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118,* 523-551.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28,* 5623-5630.

Jacobs, R.A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1,* 295-307.

McClure, S.M., Berns, G.S., & Montague, P.R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38,* 339-346.

Otto, A. R., & Love, B. C. (2010). You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making. *Judgment and Decision Making, 5*(1), 1–10.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., & Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behavior in humans. *Nature, 442,* 2006.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience, 9*(7), 545–556.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.) *Classical conditioning II: Current research and theory.* New York: Appleton-Century-Crofts.

Rumelhart, D.E., McClelland, J.E. & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vols. 1 and 2. Cambridge, MA: MIT Press.

Samanez-Larkin, G.R., Worthy, D.A., Mata, R., McClure, S.M., & Knutson, B. (2014). *Cognitive, Affective, & Behavioral Neuroscience, 14,* 672-682.

Schultz, W., Dayan, P., & Montague, P.R. (1997). A neural substrate of prediction and reward. *Science, 275, 1593-1598*

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23,* 473-500.

St-Amand, D., Sheldon, S., & Otto, A. R. (in press). Modulating episodic memory alters risk preference during decision-making. *Journal of Cognitive Neuroscience*.

Steingroever, H., Wetzels, R., & Wagenmakers, E.J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision, 1,* 161-183.

Stewart, N., Chater, N., & Brown, G.D.A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1-26.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779-804.

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV,* 96-104.

Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8,* 229-256.

Worthy, D.A., Maddox, W.T., & Markman, A.B. (2007). Regulatory fit effects in a choice task. *Psychonomic Bulletin & Review, 14,* 1125-1132.

Worthy, D.A., & Maddox, W.T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology, 59,* 41-49.

Worthy, D.A., Pang, B., & Byrne, K.A. (2013). Decomposing the roles of perseveration and expected value in models of the Iowa gambling task. *Frontiers in Psychology, 4,* 640.

Yechiam, E., & Busemeyer, J.R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision-making. *Psychonomic Bulletin & Review, 12,* 387-402.

Yechiam, E. & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology, 51,* 75-84.