Cognition 193 (2019) 104042

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Learning reward frequency over reward probability: A tale of two learning rules

Hilary J. Don[a,*], A. Ross Otto[b], Astin C. Cornwall[a], Tyler Davis[c], Darrell A. Worthy[a]

[a] Texas A&M University, United States
[b] McGill University, Canada
[c] Texas Tech University, United States

ARTICLE INFO

Keywords:
Reinforcement learning
Delta rule
Decay rule
Reward frequency
Prediction error
Probability learning

ABSTRACT

Learning about the expected value of choice alternatives associated with reward is critical for adaptive behavior. Although human choice preferences are affected by the presentation frequency of reward-related alternatives, this may not be captured by some dominant models of value learning, such as the delta rule. In this study, we examined whether reward learning is driven more by learning the probability of reward provided by each option, or how frequently each option has been rewarded, and assess how well models based on average reward (e.g. the delta model) and models based on cumulative reward (e.g. the decay model) can account for choice preferences. In a binary-outcome choice task, participants selected between pairs of options that had reward probabilities of 0.65 (A) versus 0.35 (B) or 0.75 (C) versus 0.25 (D). Crucially, during training there were twice the number of AB trials as CD trials, such that option A was associated with higher cumulative reward, while option C gave higher average reward. Participants then decided between novel combinations of options (e.g., AC). Most participants preferred option A over C, a result predicted by the Decay model, but not the Delta model. We also compared the Delta and Decay models to both more simplified as well as more complex models that assumed additional mechanisms, such as representation of uncertainty. Overall, models that assume learning about cumulative reward provided the best account of the data.

1. Introduction

How do we take into account the amount of experience we have with choice alternatives when making decisions? For example, imagine you are deciding whether to go to an old restaurant that you have visited frequently, or a relatively new restaurant that you have visited only a few times. Your choice could be based on the average quality of each restaurant, for example, the food, service, and atmosphere. Alternatively, your choice could be based on the cumulative number of positive experiences you have had with each restaurant. Since you have had much more experience with the old restaurant, this may then bias you towards choosing it, even if the average quality of the new restaurant has been higher (cf. Bornstein & D'Agostino, 1992).

Such decision processes are typically studied in laboratory settings using reinforcement learning tasks in which participants learn the relationship between alternative actions and subsequent rewards through experience. The expected value – an estimate of future reward – for each alternative is learned through trial-and-error, and options with a higher expected value are more likely to be chosen in the future. Most prominent models of reinforcement learning assume that expected value is based on the average reward provided by each option. The *Delta rule* (Rescorla & Wagner, 1972; Widrow & Hoff, 1960; Williams, 1992), for example, is one of the most commonly used learning rules across domains, including reward and value learning in decision-making, category learning, and associative learning paradigms (e.g. Busemeyer & Stout, 2002; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Gluck & Bower, 1988; Jacobs, 1988; Rumelhart & McClelland, 1986; Sutton & Barto, 1981; 1998). This rule updates expected values by learning about average reward probability, such that the frequency with which each option is experienced will not affect its value.

However, previous work suggests that the way people learn an estimate of the expected value of an option may not be as simple as the probability of that option yielding a reward. Estes (1976) manipulated the frequency of choice options, and found that probability judgments were heavily influenced by how frequently each option had been encountered. Accordingly, Estes suggested that people are more likely to translate memories of rewarded events associated with each alternative into probability judgments, rather than represent such probabilities

* Corresponding author at: Department of Psychological & Brain Sciences, Texas A&M University, 4235 TAMU, 77843-4235, United States.
E-mail address: hilary.don@tamu.edu (H.J. Don).

https://doi.org/10.1016/j.cognition.2019.104042
Received 7 November 2018; Received in revised form 4 August 2019; Accepted 6 August 2019

directly (cf. Murty, FeldmanHall, Hunter, Phelps, & Davachi, 2016). Typically, the same options that provide greater rewards on average are also associated with the highest cumulative reward. This is because in the majority of studies of choice behavior all options are presented with the same frequency. However, outside of an experimental context, choice options are generally not experienced or encountered in equal frequency. For example, grocery stores might provide some items year-round, but other items only seasonally, such that people will have much more experience with the common items. Model-based predictions about the way people behave in these situations may differ depending on whether a given model assumes that people learn about average or cumulative reward associated with each choice alternative.

Intriguingly, this issue has been overlooked by dominant learning models which are used to characterize experience-based decision-making. Learning rules are a key component of formal models of cognition. They dictate how models acquire, update, and maintain information about the values of choice alternatives, the weights between network connections, and the strengths of memories. The way in which learning rules are formulated can therefore have substantial effects on the types of information a model is able to learn, and thus what assumptions cognitive theories explicitly or implicitly make about the mechanisms underlying cognition. Despite their central importance to models of choice, there has been little work to systematically compare learning rules in their sensitivities to different types of information. Thus, our aim for this study was to determine whether expected value is driven by the average reward or cumulative reward yielded by choice alternatives, by directly comparing learning rules that assume either average or cumulative estimates of expected value. Specifically, we compared these learning rules to human choice preferences when the frequency and probability of reward provided by alternative options differed.

We first designed a reinforcement learning task that differentiates between learning expected value based on average reward and learning expected value based on cumulative reward. Model performance was then simulated *ex ante* on this task across a large range of parameter values to verify the models make diverging predictions. We then collected human data on the reinforcement learning task and fit each model to the data to obtain best fitting parameter values and compare model fits. The models were then simulated *ex post* using the best-fitting parameter values to determine whether the models can reproduce the behavioral effects of interest, thereby providing conditions for falsifiability of the models in question (see Palminteri, Wyart, & Koechlin, 2017). Finally, we ran a cross-fitting procedure to assess model recovery.

1.1. Reinforcement learning task

To test learning rule model predictions, we used a reinforcement learning task that dissociates reward frequency from reward probability. In this task, participants selected between two options on each trial, and received either reward or no reward, based on fixed probabilities tied to each option. On some trials, participants learned to choose between option A, rewarded 65% of the time and option B, rewarded 35% of the time. On other trials, participants learned to choose between option C, rewarded 75% of the time, and option D, rewarded 25% of the time. Critically, participants were given 100 AB trials, but only 50 CD trials. This should create a situation where option A is associated with the most cumulative reward, whereas option C has provided the most reward on average. Upon test, participants were presented with several different combinations of the choice options, each presented 25 times. The key comparison occurs on CA trials, as models based on average reward should predict more C choices (higher average reward), while models based on cumulative reward should predict more A choices (higher cumulative reward). Further details about the task will be presented below.

2. *Ex ante* simulations

2.1. Basic models

To verify our predictions for models based on average and cumulative reward, we first simulated the task with two models where expected value (EV) is based either exclusively on the average reward provided by each option, or exclusively on the cumulative reward provided by each option. For the basic average model, if rewards (r) are coded as 1 when a reward is given and 0 when a reward is not given then the cumulative reward value (CRV_j) for each j option is computed on each t trial as:

$$CRV_j(t+1) = CRV_j(t) + r(t)\cdot I_j \tag{1}$$

where I_j is simply an indicator value that is set to 1 if option j is selected on trial t, and 0 otherwise. The number of times each j option has been selected (N_j) is used as the denominator, and expected values (EV) are simply the average reward values:

$$EV_j(t+1) = \frac{CRV_j(t+1)}{N_j} \tag{2}$$

In the basic cumulative model, expected values for each option are simply the cumulative reward values from Eq. (1).

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

$$P\,|C_j(t)| = \frac{e^{\beta\cdot EV_j(t)}}{\sum_1^{N(j)} e^{\beta\cdot EV_j(t)}} \tag{3}$$

where $\beta = 3^c - 1 (0 \le c \le 5)$, and c is a log 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. Defining β in this way allows it to take on a very large range of values (0–242), and is equivalent to setting a prior on beta with a truncated exponential distribution.[1]

2.2. Reinforcement learning models

Delta and Decay rule models also base expected value on average and cumulative reward, respectively, but include recency parameters that allow the models to better approximate human behavior. Our original goal of the study was to compare Delta and Decay models because the models are not overly flexible, which allows for strong inference since each model can be excluded or falsified if participants exhibit certain patterns of behavior that each model cannot predict 'ex ante' (Busemeyer & Wang, 2000; Platt, 1964; Roberts & Pashler, 2000). However, we also considered some more complex variants of these models, which are presented below.

2.2.1. Delta rule model

The Delta rule (Rescorla & Wagner, 1972; Widrow & Hoff, 1960; Williams, 1992) updates expected values based on prediction error, that is, the difference between what was expected and what was received in a given instance. Expected values will therefore approximate the

[1] In the fits and simulations presented in this paper, EVs were scaled by subtracting the minimum EV from each EV such that the difference among EVs was preserved, but the values were scaled in a way that worked better with the computer program we used (R), without producing "NA" from the exponential function. The softmax rule evaluates the numerical distance between EVs to determine action selection probability so shifting values in this way did not affect the models' predictions (Worthy, Maddox, & Markman, 2008). Additionally, the maximum value entered into the exponential function for the softmax rule was 700, as values greater than this lead to "NA" in most computer programs, including R. Full analysis code is available at: https://osf.io/v57wf/.

recency-weighted average reward associated with each option, such that the frequency with which each option is experienced will not affect its value. Expected values in the delta rule model are calculated as:

$$EV_j(t+1) = EV_j(t) + \alpha \cdot (r(t) - EV_j(t)) \cdot I_j \quad (4)$$

The update function on the delta rule means that expected values are only updated for the chosen alternative on each trial. If participants choose A for an AB pair, they update their information about A, but not B. The portion of Eq. (1) in parentheses is known as the prediction error, and it is modulated by the learning rate parameter ($0 \leq \alpha \leq 1$). Higher values of α indicate greater weight to recent outcomes, while lower values indicate less weight to recent outcomes. When $\alpha = 0$ no learning takes place and expected values remain at their starting points, and when $\alpha = 1$ expected values are equal to the last outcome received for each option. The probability of choosing each alternative was determined by entering EVs into the Softmax rule in Eq. (3).

2.2.2. Decay rule model

The Decay rule (Erev & Roth, 1998; Yechiam & Busemeyer, 2005; Yechiam & Ert, 2007) is a less prevalent learning rule than the Delta rule, and updates expected values based on cumulative instances of reward associated with each option. Rather than updating expected values on the basis of prediction error, value estimates are dictated by how often (in total) each option has yielded reward in the past. In turn, options that have been encountered more frequently should be associated with more reward overall, and should receive greater value. The Decay rule assumes that outcomes are stored directly in memory, but this memory trace decays over time. Specifically, on each t trial the EV for each j option is updated according to:

$$EV_j(t+1) = EV_j(t) \cdot (1 - A) + r(t) \cdot I_j \quad (5)$$

where A is a decay parameter,[2] and $r(t)$ is the reward given on each trial. As in Eq. (1), I_j is an indicator variable that is set to 1 if option j was selected on trial t, and 0 otherwise, such that EVs will increment by the reward for the chosen option only, but all options will decay on every trial. Again, the probability of choosing each alternative was determined by the Softmax rule (Eq. (3)).

In summary, the key difference between the Delta and Decay rules is that the Delta rule will learn the *average* reward provided by each option, modulated by a learning rate parameter, while the Decay rule will learn a decaying representation of the *cumulative* reward associated with each option.

2.3. Method

We simulated data sets across the parameter space by systematically incrementing values of c, along with either α or A for the Delta and Decay models. α or A varied from 0 to 1 in increments of 0.05, and c varied from 0 to 5 in increments of 0.25. For each parameter combination, we performed 1000 simulations where each model learned the training phase of the task. After the training phase, we then took each model's predicted probability of choosing C on the critical CA trials (Eq. (3)), and these probabilities were averaged across the 1000 simulations for each parameter combination.

2.4. Results & discussion

Averaged across c parameters, the average model predicted a greater proportion of C choices on CA trials (mean p(C) = 0.78), while the cumulative model predicted greater A choices (mean p(C) = 0.24).

[2] We used (1-A) as the decay parameter so that higher values indicate more decay and lower values indicate less decay, such that they are more comparable to the learning rate parameter in the Delta model. In both models higher values indicate a greater reliance on recent outcomes.

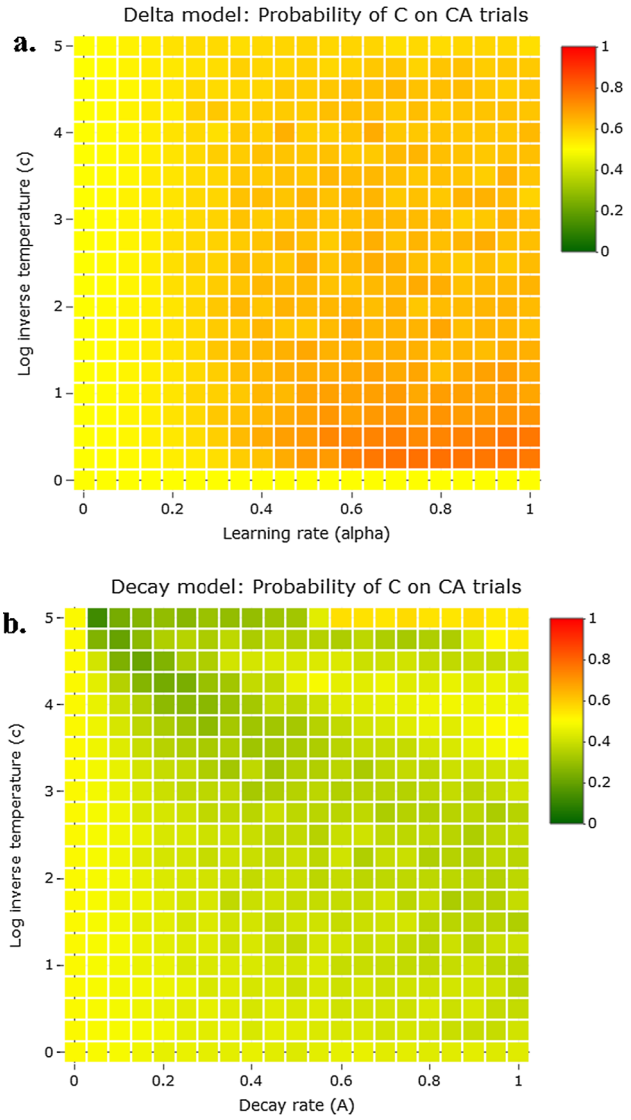

Fig. 1. The average probability of selecting C on CA trials across 1000 simulations for each parameter combination from (a) the Delta model and (b) the Decay model.

Fig. 1 plots the average probability of selecting option C on CA trials for each parameter combination for the Delta (a) and Decay (b) models. Delta model predictions range from 0.5 to about 0.75, while Decay model predictions range from about 0.15 to 0.55. Overall, the Delta model generally predicts more C choices (mean p(C) = 0.60), while the Decay model generally predicts more A choices (mean p(C) = 0.41). Thus, this task is useful for evaluating whether human behavior is more in line with model predictions that expected value is based on average or cumulative reward, respectively. We therefore tested how well these predictions align with human behavior.

3. Choice experiments

To summarize, the Delta rule model assumes that people learn average reward or average reward probabilities, and that the frequency with which each alternative has been encountered should not affect its value. The Decay rule assumes that people learn a cumulative representation of reward associated with each option, and that options that have been more frequently rewarded should hold higher value.

To test these predictions, we conducted three choice experiments, which were all variants of the same basic task design. The critical

components of the task remained consistent across experiments so we present them together for concision. We were most interested in choices made on CA test trials. From the simulations above it is clear that the models based on average reward predict more C choices, while models based on cumulative reward predict more A choices. We evaluated choices on these trials and performed fits of each model to the choice data to evaluate which model best characterizes the observed human behavior.

3.1. Method

3.1.1. Participants

One-hundred and thirty-three participants from Texas A&M University (88 female, mean age = 19.1, *SD* = 1.16) participated in the experiment for partial fulfillment of a course requirement. The Internal Review Board approved the experiment, and participants gave informed consent. Sample sizes for Experiments 1, 2, and 3 were 33, 50, and 50, respectively. Experiment 1 was completed near the end of a term, with the goal of running participants up until that point (see Witt, Donnellan, & Orlando, 2011, for subject pool effects as a function of time of semester). The sample size of 50 for the last two experiments was determined semi-arbitrarily, and as a value with enough power to conduct one-sample t-tests for choice proportions on test trials against specific expected choice proportions (detailed below). For the critical comparison of choice proportions on CA trials, with 50 participants we had about 0.96 power to detect an effect, assuming a moderate effect size of d = 0.50.

3.1.2. Materials and procedure

Participants performed the experiment on PCs using Matlab software with Psychtoolbox-2.54. Participants first completed several questionnaires that are described in the Supplementary Materials. The reward structure was identical to the task described in the introduction, and is detailed in Table 1. Fig. 2 shows example trial sequences from the different versions of the task. Participants were told that they would make repeated choices on each trial and that they would either receive a reward of one point or zero points on each trial. During the first 150 trials of the task participants made choices between options A versus B, or C versus D. There were 100 AB trials, and 50 CD trials, with each trial type randomly distributed over the 150 training trials. The reward probabilities associated with options A-D were [0.65 0.35 0.75 0.25]. The option order was the same in Experiment 1 for all participants, but counterbalanced in Experiment 2 and Experiment 3 because we realized that keeping the options the same might bias participants' choices.[3] In Experiments 2 and 3 AB and CD pairs were presented on either the left or the right, counterbalanced across participants, and the location of options within each pair was further counterbalanced such that their order could be AB, BA, CD, or DC. For each participant, the option locations were kept consistent throughout both training and test.

Upon selecting an option, a random number from 0 to 1 was drawn from a uniform distribution in the background of the computer program. If the number drawn was less than the probability of reward associated with each option then participants received one point, otherwise they received zero points.

After 150 training trials participants were told that they would now make choices between different pairs of options. They then performed 100 additional trials with pairs CA, CB, AD, and BD. There were 25 trials of each type randomly distributed across test trials. In Experiments 1 and 2, participants received feedback in the test phase as well, but in Experiment 3 this phase was performed without feedback.

[3] Specifically, in Experiment 1 options A and C were always presented on the left on AB and CD trials, and option A was presented on the left during CA trials. This may have biased participants toward selecting A on CA trials because 'left' had been the most rewarding response for both AB and CD trials during training.

Table 1
Task design.

Phase	Number of trials	Choice options (p(reward))
Training	100	A (0.65) vs. B (0.35)
	50	C (0.75) vs. D (0.25)
Test	25	A (0.65) vs. C (0.75)
	25	A (0.65) vs. D (0.25)
	25	B (0.35) vs. C (0.75)
	25	B (0.35) vs. D (0.25)

In Experiment 2 and 3, participants also completed a final 50-trial phase where they could select from any of the four options. The results of this phase, which are of no specific interest to the present study, are presented in the Supplementary Material. Participants were not given monetary bonuses in Experiment 1, and only told to earn as many points as possible. In Experiment 2 and 3, participants were offered a monetary bonus of $0.10 USD for each point they received in the 100 trial test phase and the final 50-trial four-choice phase.

3.1.3. Data analysis

Below we compare the proportion of each objectively optimal choice for each trial type during the learning and test phases of the tasks. For AB and CD trials we conduct one-sample t-tests against chance, 0.50, to determine if, overall, participants learned that A and C were the best alternatives within each pair. For the test phase we perform one sample t-tests with the objective reward ratio between the two alternatives as the comparison metric. Thus for CA trials we compare against: 0.75/(0.75 + 0.65) = 0.5357, and for CB, AD and BD trials against values of 0.6818, 0.7222, and 0.5833, respectively. We also report a comparison for CA trials against a value of 0.50, as well as comparisons between each experiment to evaluate whether the results were consistent. The reason for using the objective reward ratio is to examine whether participants' behavior differed from what would be expected from probability matching if participants had accurate knowledge of the probability of reward associated with each option. This is different than simply examining for departures from chance or random behavior by using a test-metric of 0.5. Thus using both the reward ratios and 0.5 as test metrics allows us to test two separate hypotheses: was behavior different than predicted from accurate knowledge of reward probabilities for each option, and was behavior different than expected from chance?

We report the t and *p*-values from each test, as well as Bayes Factors in favor of the alternative hypothesis (BF_{10}), in this case that the observed proportion of optimal choice selections differs from the value expected based on the objective reward ratio between the two options. Bayes Factors were computed in JASP (jasp-stats.org) using the default priors. We consider a Bayes Factor of 3 or more to be analogous to a critical threshold, although Bayes Factors can be interpreted continuously as the odds in favor of the alternative hypothesis (Wagenmakers et al., 2018). Bayes Factors less than 1 indicate more support for the null than the alternative hypothesis, and a Bayes Factor less than 1/3 would suggest moderate support for the null hypothesis (analogous to a BF_{10} of 3 in favor of the alternative hypothesis).

We also fit each model to participants' learning and test data individually. We compared the model fits using the Bayesian Information Criterion (BIC; Schwarz, 1978), and examined the degree to which the Decay model fit the data better than the Delta model by computing: $BIC_{Delta\text{-}Decay} = BIC_{Delta} - BIC_{Decay}$. This BIC difference can then be transformed into a Bayes Factor representing the evidence that the Decay rule is the better model: $BF_{10\text{-}Decay} = \exp(BIC_{Delta\text{-}Decay}/2)$ (Wagenmakers, 2007).

3.1.4. Data availability

Data and experiment and analysis code are available on the Open Science Framework: https://osf.io/v57wf/.

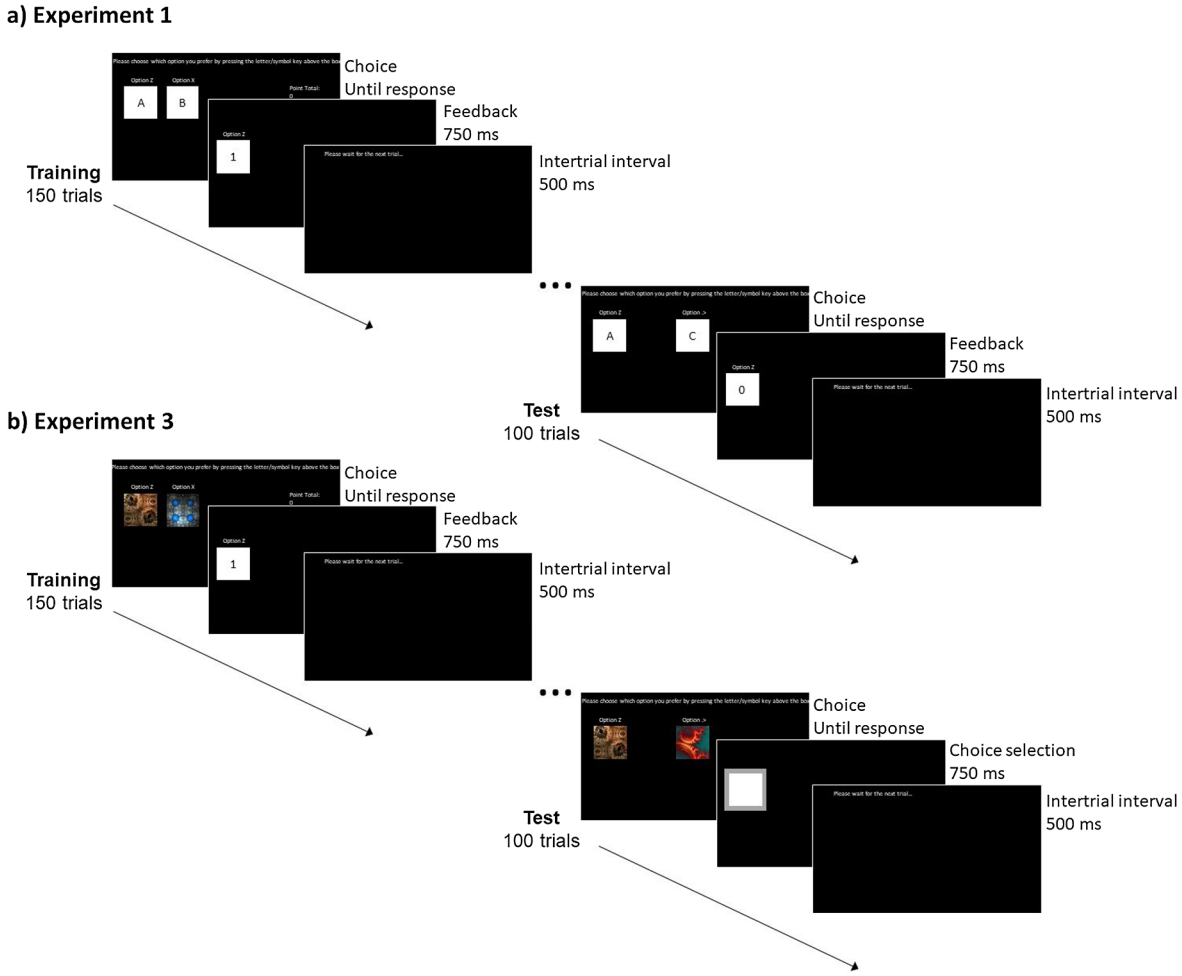

Fig. 2. Example trial sequences of the training and test phase from (a) Experiment 1 (with feedback at test), and (b) Experiment 3 (without feedback at test).

3.2. Results

3.2.1. Training trials

Fig. 3 shows the proportion of optimal choices across training, divided into six blocks of 25 trials for each experiment. Figures for all data collapsed across experiments is shown in the Supplemental Material. A repeated measures analysis of variance (ANOVA) with block (1–6) and trial type (AB vs. CD) as factors showed a significant linear effect of block, $F(1,132) = 44.57$, $p < .001$, $\eta_p^2 = 0.252$, indicating learning across the training phase. There was also a significant quadratic effect of block, suggesting learning is reaching asymptote towards the end of training, $F(1,132) = 23.19$, $p < .001$, $\eta_p^2 = 0.149$. However, there was no main effect of trial type, $F(1,132) = 0.002$, $p = .966$, $BF_{10} = 0.055$, interaction between the linear effect of block and trial type, $F(1,132) = 0.249$, $p = .618$, $\eta_p^2 = 0.002$, or interaction between the quadratic effect of block and trial type, $F(1,132) = 1.56$, $p = .213$, $\eta_p^2 = 0.012$. Overall, A and C choices were both well above chance, A choices: $M = 0.663$, SD = 0.195, $t(132) = 9.62$, $p < .001$, $BF_{10} > 1000$; C choices: $M = 0.662$, SD = 0.209, $t(132) = 8.93$, $p < .001$, $BF_{10} > 1000$. There were no significant differences in optimal choices between experiments on AB trials, $F(2,130) = 2.42$, $p = .093$, $BF_{10} = 0.54$, or CD trials, $F(2,130) = 2.96$, $p = .056$, $BF_{10} = 0.88$.

3.2.2. CA test trials

The critical CA test trials determine whether participants prefer the option with higher average reward, or higher reward frequency. Fig. 4a shows the average proportion of C choices on these trials. On average, participants selected C less often than would be expected from the objective reward ratio, $M = 0.430$, SD = 0.289, $t(132) = -4.22$, $p < .001$, $BF_{10} = 347.4$, and the median for C choices was 0.40. We also conducted a one-sample *t*-test using a test value of 0.50. This provided moderate evidence that the proportion of C choices was lower than would be expected from chance, $t(132) = -2.79$, $p = .006$, $BF_{10} = 3.96$. Fig. 3 plots choices on CA trials across training. A repeated measures ANOVA with block (1–4) and feedback (with feedback vs. without feedback) as factors showed no linear effect of block, $F(1,131) = 1.10$, $p = .296$, $\eta_p^2 = 0.008$, and no interaction between feedback and the linear effect of block, $F(1,131) = 0.60$, $p = .440$, $\eta_p^2 = 0.005$. Separate ANOVAs for each feedback condition also showed no linear effect of block in experiments with feedback at test, $F(1,82) = 0.049$, $p = .825$, $\eta_p^2 = 0.001$ or without feedback at test, $F(1,49) = 1.37$, $p = .248$, $\eta_p^2 = 0.027$. Plots of choice preferences across the test phase for each experiment separately can be found in the Supplemental Material. Fig. 4b plots of the distribution of C choices on CA trials for each experiment. Visual inspection of these plots suggests some bimodality, with some participants clearly preferring option C, but more showing a strong preference toward option A. The plot for Experiment 3 indicates the greatest extent of bimodality, possibly due to the lack of feedback during the test phase in Experiment 3; a small subset of participants strongly preferred option C, but most showed a bias toward option A.

3.2.3. Remaining test trials

On CB test trials, participants chose C significantly less than the objective reward ratio of 0.6818, $M = 0.519$, SD = 0.318, $t(132) = -5.93$, $p < .001$, $BF_{10} > 1000$. The median value was 0.48. Even though option C was 40% more likely to give a reward than option

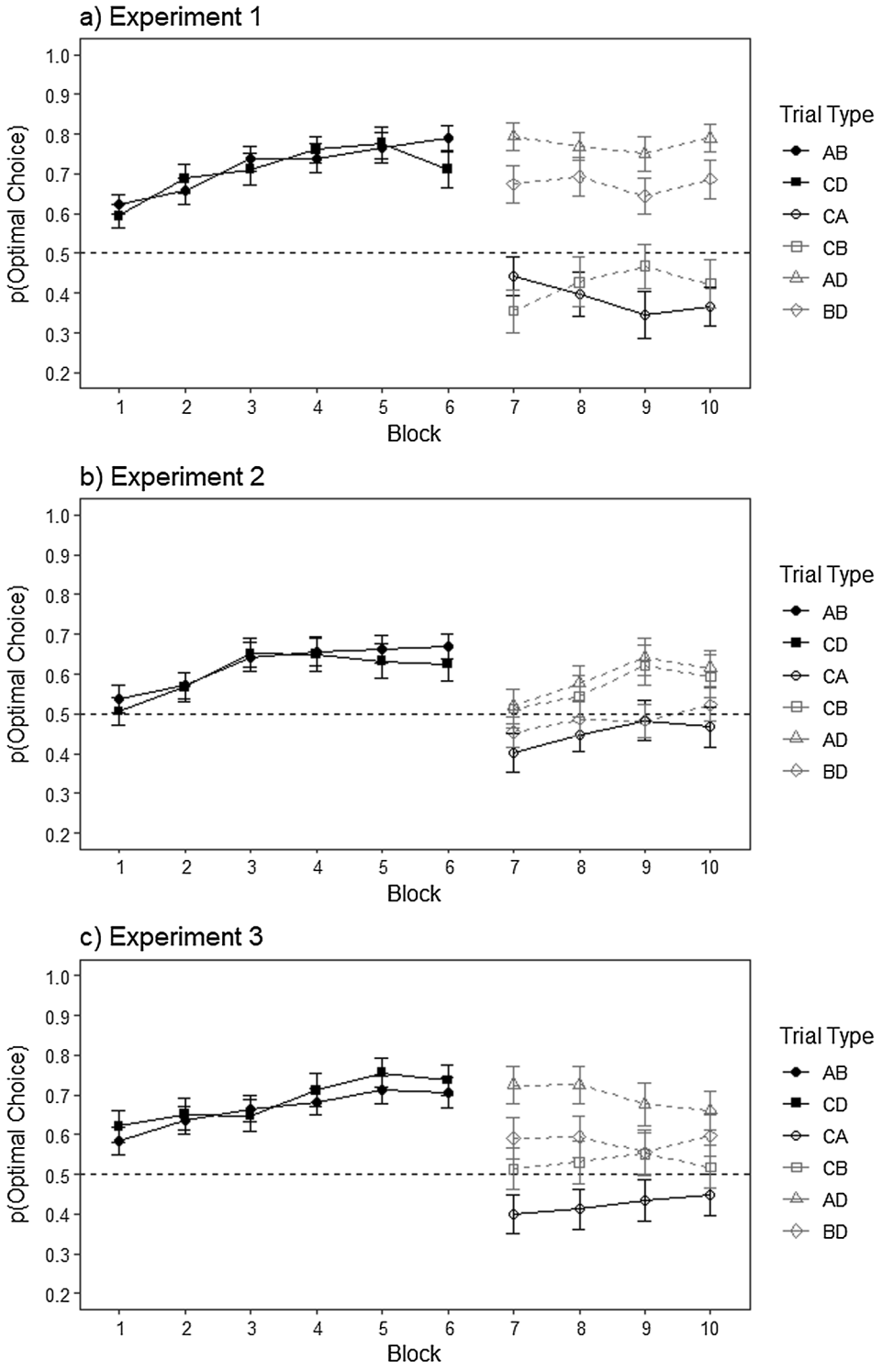

Fig. 3. Proportion of optimal choices for AB and CD trials across training and the four test trials in the test phase, split across blocks of 25 trials in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3.

B, many participants did not show a strong preference for option C on CB trials. For AD and BD trials, participants selected the better alternative in accordance with their objective reward ratios. The proportion of A choices on AD trials did not differ from the reward ratio of 0.7222, $M = 0.681$, SD = 0.275, $t(132) = -1.73$, $p = .086$, $BF_{10} = 0.407$. Similarly, the proportion of B choices on BD trials did not differ from the reward ratio of 0.5833, $M = 0.569$, SD = 0.281, $t(132) = -0.598$, $p = .551$, $BF_{10} = 0.115$. Thus, when the most frequently presented items were dominant within a pair, choice probabilities were similar to the objective reward ratios, or probability matching, but when a less frequently presented item was dominant within a pair, choice proportions were more consistent with how frequently the item had been presented.

4. Model comparisons

We first focus on comparing the fits of the Delta and Decay models to the choice data. These are basic level models that are neither as

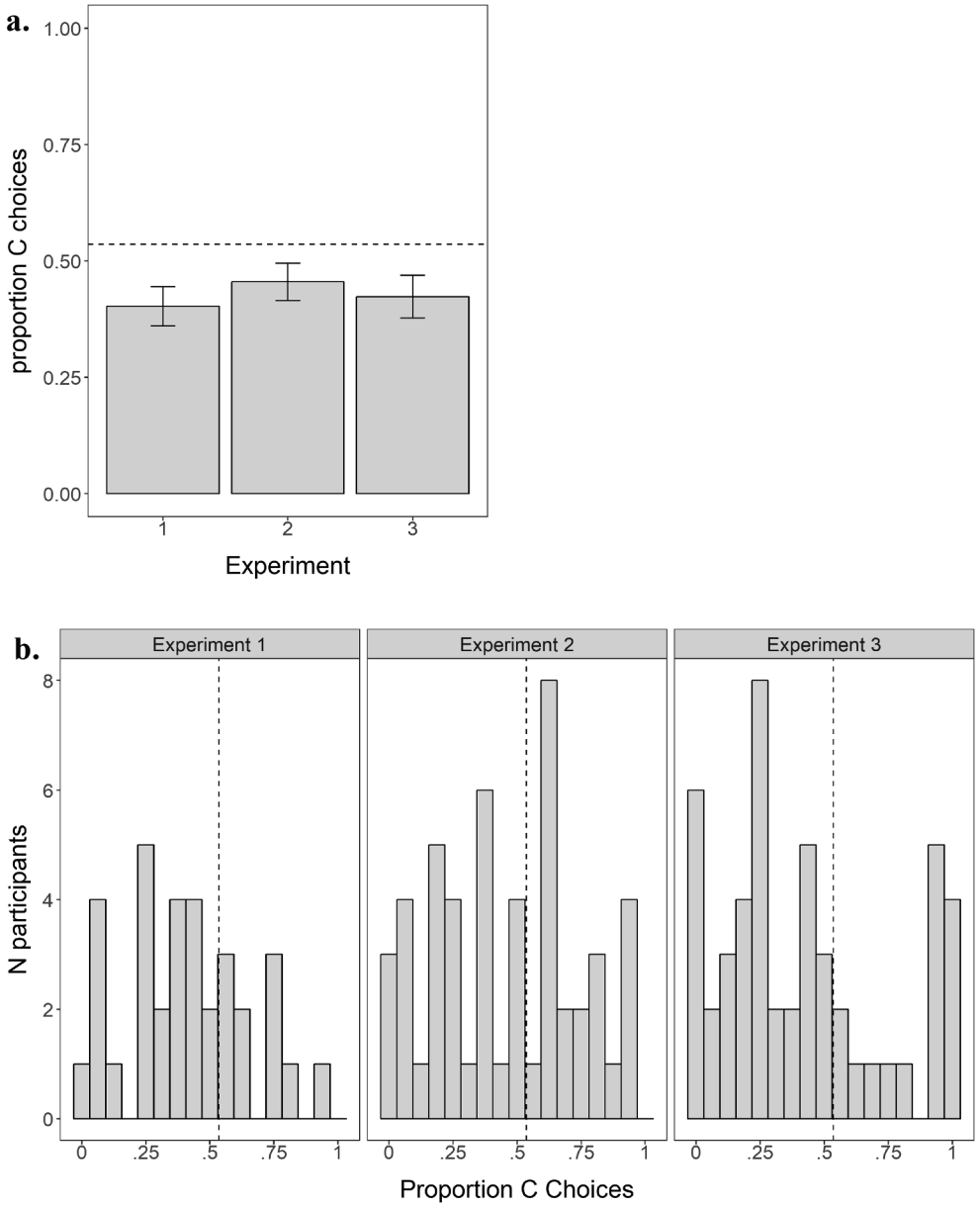

Fig. 4. C selections on CA test trials. (a) shows average proportion of C selections for each experiment. (b) shows the frequency distributions of C choices on CA test trials where Experiments 1–3 are presented from left to right. Error bars represent standard errors of the mean. Dashed lines represent the objective reward ratio.

simple as the average and cumulative models, nor as complex as the additional models we will present below. As previously noted, our focus was on comparing the Delta and Decay models as their limited flexibility allows us to falsify models if participants' behavior is qualitatively different than that predicted by the models (Roberts & Pashler, 2000). Next, we present the more complex models along with comparisons to the Decay model, as well as the simpler Delta model nested with each more complex model, then we present a section comparing all models together.

4.1. Delta vs. Decay models

To compare the models directly, we fit each model individually to each subject's training and test phase data by maximizing the likelihood of each model's next step ahead predictions. The Decay model advantage was taken as $\Delta BIC_{Delta\text{-}Decay}$. Fig. 5a plots the average Decay model advantage for each experiment. The average BIC difference was 22.57 which corresponds to a Bayes Factor of over 7.9×10^4 in favor of the Decay model (Wagenmakers, 2007). 72.22% of participants were best fit by the Decay model; a binomial test suggests this difference is well above 50% expected by chance, $p < .001$, $BF_{10} > 1000$.

We also compared the Delta and Decay models to a random or null model by computing McFadden's pseudo R^2 (McFadden, 1973). Both models fit better, overall, than the random model. For the Delta rule model pseudo R^2 values for Experiment 1–3 were 0.142, 0.134, and 0.178. For the Decay rule model pseudo R^2 values were 0.239, 0.195, and 0.226. Thus the Decay model explained about 7% more of the variance in behavior than the Delta model, although a great deal was still left unexplained by either model. To illustrate the model fits on an individual level, pseudo R^2 was also calculated for each participant, which are plotted in Fig. 5b.

4.2. Extended delta models

The results suggest that expected value is more likely to be based on cumulative reward than average reward, in support of the Decay model

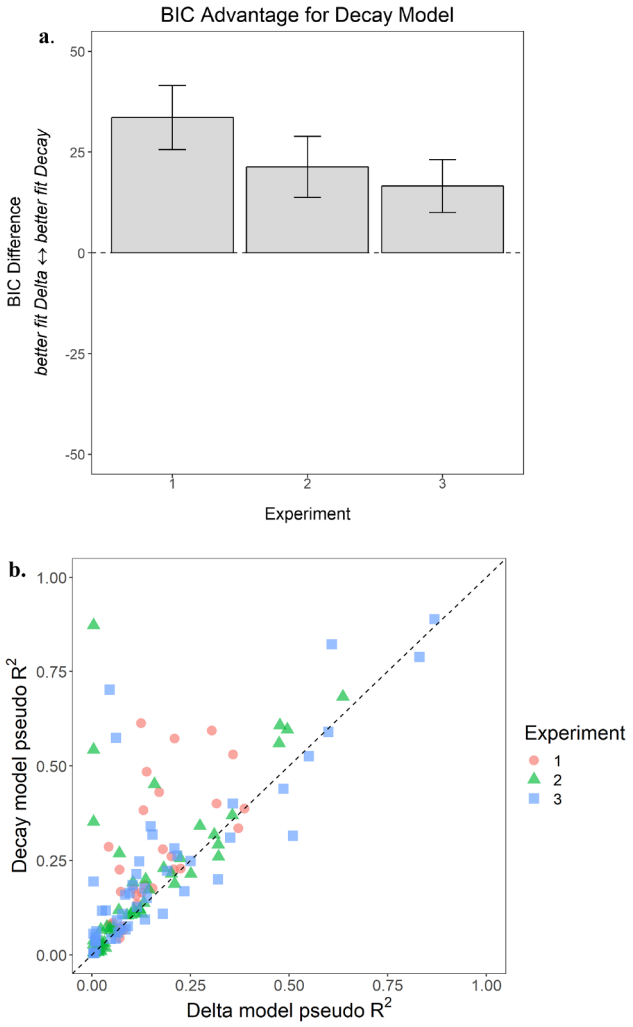

Fig. 5. Model comparisons including (a) average BIC advantage for the Decay model ($BIC_{Delta\text{-}Decay}$), where a higher score indicates a greater advantage for the Decay model, and error bars represent standard errors of the mean, and (b) McFadden's pseudo R^2 for each participant for the Delta and Decay models, where higher scores indicate greater support for each model over the null model.

predictions. However, it is possible that additional parameters may allow the Delta model to account for biases towards options with higher cumulative value. We will first compare these models to the nested Delta and Decay models, and then make comparisons across all models.

4.2.1. Delta with decay

The decay model differs from the Delta model not only in learning cumulative rewards, but also in the addition of a decay, or forgetting rate parameter. A Delta model with an additional decay rate parameter may therefore account for frequency effects in a similar way to the Decay model. This is similar to previous models that have added decay or forgetting parameters to delta-learning models (e.g. Collins & Frank, 2012; Worthy & Maddox, 2014). That is, as C is experienced and therefore chosen less often than A, its value may decay to a greater extent than A. We therefore fit the data to a delta model that included a decay rate parameter on all choice options:

$$EV_j(t+1) = EV_j(t)\cdot A + \alpha\cdot(r(t) - EV_j(t))\cdot I_j \quad (6)$$

The delta with decay model provided a significantly better fit to the data than the delta model ($\Delta BIC_{delta\text{-}deltaDecay} = 19.60$, $BF = 1.8 \times 10^4$), but did not fit better than the Decay model ($\Delta BIC_{Decay\text{-}deltaDecay} = -2.96$, $BF = 0.23$).

4.2.2. Delta with ± α model

There is evidence that learning rates for positive and negative prediction errors may differ (Christakou et al., 2013; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; St-Amand, Sheldon, & Otto, 2018). Specifically, Lefebvre and colleagues found that their participants exhibited an "optimism bias" where positive reward prediction errors were weighted more heavily than negative reward prediction errors. It is possible that including separate learning rate parameters for positive and negative prediction errors could allow the Delta model to predict a preference for A over C on the critical transfer trials if positive prediction errors are weighted more than negative prediction errors, as there would be a greater number of positive prediction errors for A:

If r(t) – EV(t) > 0,

$$EV_j(t+1) = EV_j(t) + \alpha_+\cdot(r(t) - EV_j(t))\cdot I_j \quad (7)$$

If r(t) – EV(t) < 0,

$$EV_j(t+1) = EV_j(t) + \alpha_-\cdot(r(t) - EV_j(t))\cdot I_j \quad (8)$$

This model also provides a better fit of the data than the original Delta model ($\Delta BIC_{Delta\text{-}Delta\ \pm\ \alpha} = 17.41$, $BF = 6.0 \times 10^3$), but does not provide a better fit than the Decay model ($\Delta BIC_{Decay\text{-}Delta\ \pm\ \alpha} = -5.16$, $BF = 0.08$).

4.2.3. Delta with uncertainty model

An alternative explanation for frequency effects here is that uncertainty about the less frequently experienced option C drives the preference for A. In other words, the greater uncertainty for C makes option A more appealing. Several studies have shown that people tend to be averse to ambiguous options (e.g., Ellsberg, 1961; Curley, Yates, & Abrams, 1986). We therefore also fit Delta and Decay models that incorporate uncertainty. Because our tasks involved binary outcomes, uncertainty associated with option *j*, U_j, was represented as the variance computed from the beta distribution.

The first uncertainty model calculated expected values using the delta rule from Eq. (4), presented above. The alpha (α_j) and beta (β_j) values from the beta distribution for each *j option* were simply the number of times each option was associated with either a reward (α_j) or non-reward (β_j). Alpha and beta values were initialized at 1 and then updated for the chosen option *i* following each trial according to:

If r(t) = 1,

$$\alpha_i(t+1) = \alpha_i(t) + 1 \quad (9)$$

If r(t) = 0,

$$\beta_i(t+1) = \beta_i(t) + 1 \quad (10)$$

Uncertainty (U_j) was then computed as:

$$U_j = \frac{\alpha_j\cdot\beta_j}{(\alpha_j + \beta_j)^2\cdot(\alpha_j + \beta_j + 1)} \quad (11)$$

The Q-value for each option (QV_j) was a combination of its expected value and the square root of its uncertainty, or variance from the beta distribution:

$$QV_i = EV_i + w_U\cdot\sqrt{U_j} \quad (12)$$

where w_U is the weight parameter for the uncertainty associated with each option. We allowed this parameter to range from −5 to 5. Positive values indicate a preference for options with greater uncertainty, and negative values predict a preference for options with less uncertainty (e.g. Payzan-LeNestour & Bossaerts, 2011). We predicted that best-fitting uncertainty weight parameters would be negative for most participants; indicating that most participants viewed uncertainty negatively. Finally, Q-values were entered into a Softmax rule similar to Eq.

(3) above to determine action selection probabilities for each alternative.

$$P\ |C_j(t)| = \frac{e^{\beta \cdot QV_j(t)}}{\sum_1^{N(j)} e^{\beta \cdot QV_j(t)}} \quad (13)$$

This uncertainty model provides a better fit of the data than the original delta model ($\Delta BIC_{Delta\text{-}DeltaUncertainty} = 16.46$, $BF = 3.7 \times 10^3$), but did not fit better than the Decay model ($\Delta BIC_{Decay\text{-}DeltaUncertainty} = -6.11$, $BF = 0.05$).

4.2.4. Decay with uncertainty model

We also fit a Decay model with the same uncertainty mechanisms as the Delta model with uncertainty presented above. Expected values were computed from Eq. (5) above, Eqs. (9)–(11) were used to compute uncertainty for each option, and Eqs. (12) and (13) were used to determine action selection probability for each alternative. This model fit the data fairly well, but did not clearly improve the fit compared to the Decay model ($\Delta BIC_{Decay\text{-}DecayUncertainty} = 0.22$, BF = 1.11).

4.3. Comparison of all models

For each of the original and extended models, we report several metrics to compare model fits and predictions. Table 2 shows a comparison of BIC, as well as the best fitting parameters for each model. As a comparison, Table 2 also presents corrected Akaike Information Criteria (AICc; Akaike, 1974; Burnham & Anderson, 2002), which tends to penalize additional parameters less heavily than BIC.

4.3.1. Model fits

Generally, the models based on cumulative reward provided a better fit of the data than the models based on average reward, and the additional parameters in the extended models improved the fit compared to the nested Delta model. Overall, the Decay model and Decay with uncertainty model provided the best fit of the data, followed by the extended Delta models. AICcs generally showed the same pattern of results, with the exception that the Delta with decay model fit the data just as well as the Decay model, and the Decay with uncertainty model provided a small improvement in model fit ($\Delta AICc_{Decay\text{-}DecayUncertainty} = 3.4$). For illustrative purposes, Fig. 6 shows mean EVs for each option across training for each model. The cumulative models give higher values to option A than option C by the end of training. The Delta model gives higher value to option C than option A, and this preference in value for C relative to A is reduced, but not reversed, in the extended Delta models.

Table 2
Model comparisons including average BIC, AICc, and best fitting parameter values.

				Best fitting parameters			
Model type	Model	BIC	AICc	c	α	A	w_U
Basic	Basic average	310.51	307.09	0.90			
	Basic cumulative	299.06	295.64	0.12			
Simple	Delta	304.75	298.00	1.40	0.37		
	Decay	282.18	275.44	0.44		0.19	
Extended	Delta with decay	285.15	275.19	2.60	0.25	0.87	
	Delta with ± α	287.34	277.38	1.70	α_+ = 0.31 α_- = 0.24		
	Delta with uncertainty	288.29	278.34	1.10	0.42		−1.13
	Decay with uncertainty	281.97	272.01	0.48	0.25		−0.01

4.3.2. Trial-level confusion matrices

Fig. 7 presents trial-level confusion matrices for each model, collapsed across experiments. These matrices depict the actual choices of participants on CA trials against each model's choice predictions on each trial, based on their best fitting parameters. That is, for each participant we derived the models' predictions for each presented option on each trial, based on the sequence of options and rewards that participant received. Model choice was determined probabilistically using the model probability for the participant's chosen option on that trial, calculated by the Softmax rule. If a participant chose option C on a particular trial, and the model also predicted C choice, then we would add a count to the Actual C/Predicted C cell. However, if the participant chose option C but the model predicted A, then we would add a count to the Actual C/Predicted A cell. This can be used to determine the percentage of accurate choices for each model. Considering the basic and simple models, the models based on cumulative reward tended to have more accurate predictions than the models based on average reward. Overall, including additional parameters to the Delta model slightly improved accuracy in choice predictions (0.56 for the Delta model compared to 0.57, 0.60 and 0.60 for Delta with uncertainty, Delta with decay, and Delta with ± α models, respectively). Including an uncertainty weight parameter to the Decay model did not increase accuracy above that of the original Decay model.

4.3.3. Best-fitting parameters

The best-fitting decay parameter value for the Delta with decay model was 0.87, which is higher than the Decay models' best-fitting value of 0.19. This suggests that the Delta model can provide a better fit for the data if there is a high rate of forgetting, but a high rate of forgetting is less necessary if expected values are based on cumulative reward. For the Delta ± α model, the best-fitting parameter value for α_+ was larger than that for α_-. This is consistent with the findings of Lefebvre et al. (2017), and indicates that allowing the delta model to give more weight to positive prediction errors can bias it towards choosing A more often. The Delta with uncertainty model has an average negative uncertainty weight, which equates to a preference for options with less uncertainty. This suggests that the addition of an uncertainty weight parameter may allow the Delta model to account for more A choices (see Fig. 7c and g). Indeed, the best-fitting uncertainty weight parameter was positively correlated with the proportion of C choice on CA trials, $r = 0.348$, $p < .001$. The Decay with uncertainty model's average uncertainty weight is near zero, such that it won't necessarily favor more certain choices. This is a result of some participants being best fit by a positive uncertainty weight, and some by a negative weight. It is possible that allowing uncertainty to be positively or negatively weighted adds flexibility to the model in either direction, by also allowing more C choices than the basic Decay model (Fig. 7d and h). However, for the Decay with uncertainty model, uncertainty weights were not correlated with choice preference, $r = 0.017$, $p = .847$. The addition of an uncertainty parameter likely does not improve the model fit as the Decay model naturally prefers the more frequently presented options.

5. *Ex post* simulations

As an additional test, we simulated the models using their best-fitting parameter values to determine their generative performance – how well they can reproduce the choice effects shown in the human choice data (Palminteri et al., 2017). So far, the cumulative models and extended delta models provide the best fit to the data, while the basic average and delta models provide the poorest fit. *Ex post* simulations are a different test for the models because they evaluate whether the models can reproduce the same pattern of data observed by participants from the best-fitting parameter estimates. Model fit comparisons using metrics such as BIC may allow more complex and flexible models to fit idiosyncratic patterns in the data. The added flexibility of more complex models may cause them to overfit the data, and impair their ability to reproduce the data through simulations as well as their ability to

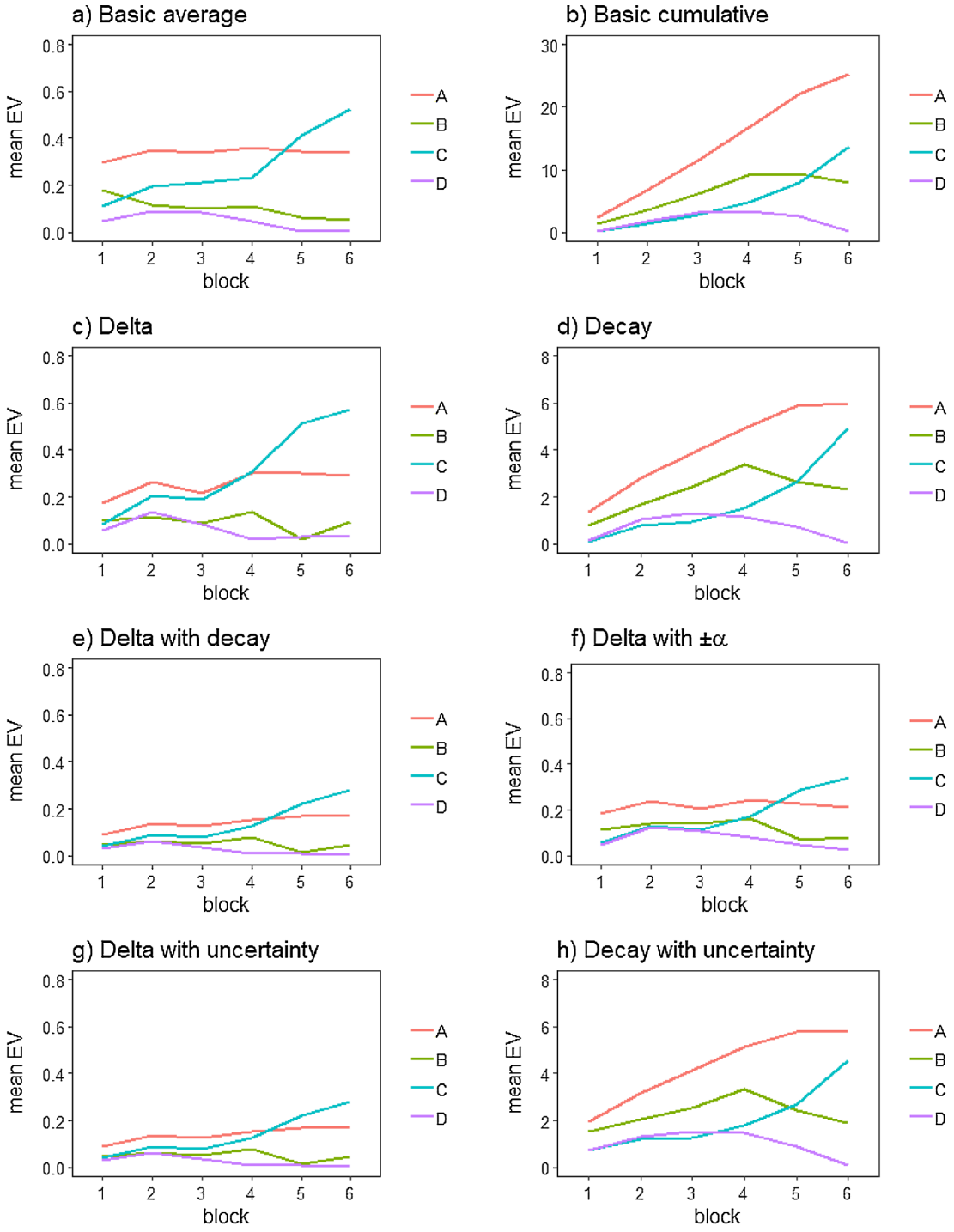

Fig. 6. Mean EVs for each option across the training phase when each model is fit to the choice data.

generalize to novel settings (Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Busemeyer & Wang, 2000). For each model, we used the best-fitting parameters for each participant when fit across the entire experiment (post-hoc simulations), and fit to the training phase only (*a priori* simulations). We were particularly interested in whether the models could reproduce performance in both the training phase and on CA trials in the test phase.

5.1. Post-hoc simulations

We ran 5000 simulations for each experiment, which each sampled one participant's best-fitting parameters for each model, fit to the entire experiment, with replacement. The predicted probability of choosing the optimal choice on each trial for each trial type was computed for each model and averaged across all 5000 simulations. These predicted probabilities of selecting the optimal choice on each trial were then compared to the observed proportion of optimal choices across participants by evaluating the root-mean-square error between predicted and observed proportions of optimal choices (RMSE; see Table 3). This method is similar to the generalization criterion method (Busemeyer & Wang, 2000), although the *a priori* simulations presented below are closer to this method because it involved testing whether best-fitting parameter estimates from a training phase can generalize by predicting behavior in a subsequent test phase. We also ran a cross-fitting

a) Basic average

	Predicted C	Predicted A	Total
Actual C	**782**	643	1425
Actual A	1028	**872**	1900
Total	1810	1515	**.50**

b) Basic cumulative

	Predicted C	Predicted A	Total
Actual C	**626**	799	1425
Actual A	564	**1336**	1900
Total	1109	2135	**.60**

c) Delta

	Predicted C	Predicted A	Total
Actual C	**843**	582	1425
Actual A	893	**1007**	1900
Total	1736	1589	**.56**

d) Decay

	Predicted C	Predicted A	Total
Actual C	**785**	640	1425
Actual A	611	**1289**	1900
Total	1396	1929	**.62**

e) Delta with decay

	Predicted C	Predicted A	Total
Actual C	**792**	633	1425
Actual A	703	**1197**	1900
Total	1495	1830	**.60**

f) Delta with ±α

	Predicted C	Predicted A	Total
Actual C	**840**	585	1425
Actual A	755	**1145**	1900
Total	1595	1730	**.60**

g) Delta with uncertainty

	Predicted C	Predicted A	Total
Actual C	**816**	609	1425
Actual A	812	**1088**	1900
Total	1628	1697	**.57**

h) Decay with uncertainty

	Predicted C	Predicted A	Total
Actual C	**782**	643	1425
Actual A	633	**1267**	1900
Total	1415	1910	**0.62**

Fig. 7. Trial-level confusions matrices for each model. Rows show the actual number of C and A choices made by participants on CA trials in all experiments, and columns show the number of predicted C and A choices made by the model with the best-fitting parameters for each participant. Grey-shaded cells indicate accurate predictions, and the percentage of accurate trial predictions is reported in the bottom right hand corner of each matrix.

procedure using these simulations to assess the capacity to recover the correct model (see Supplemental Material).

Fig. 8 shows the average model predictions for the proportion of optimal choices for AB, CD and CA trials, against participants' actual optimal choices. Figures for all training trials split by experiment are presented in the Supplemental Material. Comparing the two models of primary interest, both the Delta and Decay models provide comparable RMSEs on training trials on average, but the Decay model better approximates CA trials at test (Fig. 8c), and showed less overall error considering all trial types. On CA trials specifically, the Decay with uncertainty model provides the least error in predictions, followed by the Decay model, and the Delta models with additional parameters. However, the extended Delta models do not predict a strong choice bias towards A on CA trials.

5.2. *A priori* simulations

The *a priori* simulations sampled the best-fitting parameters when the models were fit to the training phase only, based on the generalization criterion method (Ahn et al., 2008; Busemeyer & Wang, 2000). These best-fitting parameters were then used to generate predictions for the entire data set, including the test phase, which is a stricter test of the models' performance. This method can also indicate whether more complex models over-fit the data, which would be evident if the model

Table 3
RMSE from post-hoc simulations for each model and trial type.

	Trial type						
Model	AB	CD	AC	BC	AD	BD	Average
Basic average	0.081	0.086	0.192	0.236	0.096	0.098	0.131
Basic cumulative	0.085	0.116	0.115	0.285	0.154	0.095	0.142
Delta	0.088	0.071	0.296	0.218	0.092	0.117	0.147
Decay	0.074	0.083	0.093	0.147	0.109	0.097	0.101
Delta with ± α	0.106	0.079	0.095	0.152	0.103	0.095	0.105
Delta with decay	0.113	0.104	0.098	0.126	0.120	0.108	0.111
Delta with uncertainty	0.115	0.089	0.109	0.140	0.100	0.090	0.107
Decay with uncertainty	0.081	0.078	0.076	0.129	0.087	0.084	0.089

can predict the data to which it has been fit, but not the data to which it has not been fit (Farrell & Lewandowsky, 2018). We again ran 5000 simulations of each model, randomly sampling with replacement from participants' best-fitting parameters to the training phase. RMSE for each model is shown in Table 4, and average model choice predictions across the 5000 simulations are shown in Fig. 9. See the Supplemental Material for figures for all training trials split by experiment.

The results followed a similar pattern to the post-hoc predictions. Although the Delta model made slightly better predictions than the Decay model on training trials, the Decay model again made better predictions on CA trials (Fig. 9c), and less overall error than the Delta model. On CA trials, the Decay with uncertainty model again showed the least prediction error, as well as the Decay model. The Delta models with additional parameters performed better than the basic Delta model on the test trials. However, while the Delta with decay and Delta with ± α models can predict a slight bias towards A when fit to the data (see Fig. 7e and f), they tend to predict choice close to chance in *ex post* simulations. The Delta with uncertainty model predicted C choices slightly above chance in the *a priori* simulations. The models based on cumulative reward better predict choice preferences on these trials. Thus, in these simulations, the Decay model with fewer parameters performed just as well as the more complex models on the training trials to which the models were fit, and showed less error on tests trials to which the models were not fit.

6. Model recovery

We used a cross-fitting procedure similar to that reported by Wagenmakers, Ratcliff, Gomez, and Iverson (2004) to assess the

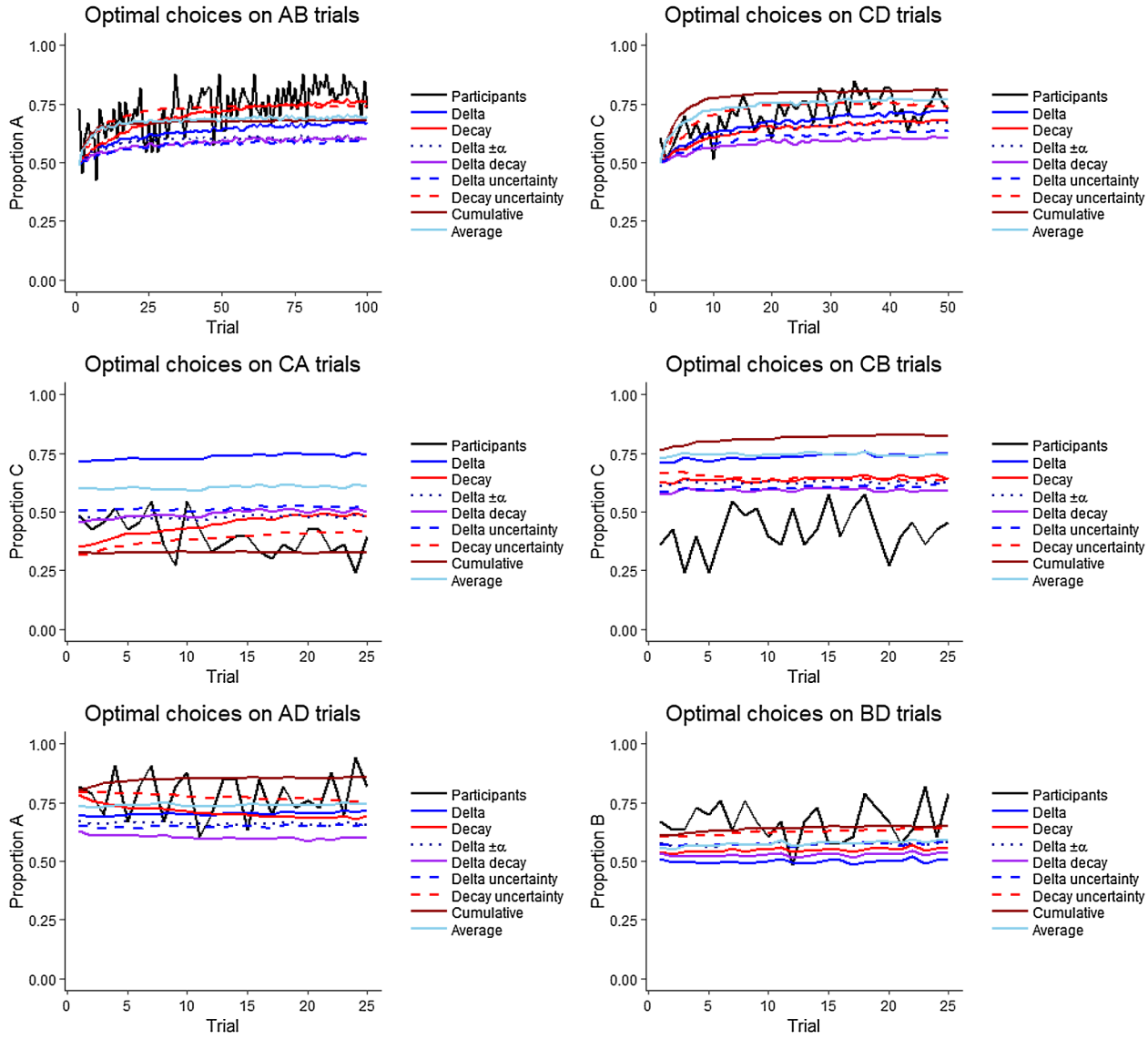

Fig. 8. Model choice predictions from post-hoc simulations against participants' actual choices shown by the solid black lines on (a) AB trials, (b) CD trials, and (c) CA trials.

Table 4
RMSE from a priori simulations for each model and trial type.

	Trial type							
Model	AB	CD	AC	BC	AD	BD	Average (training)	Average (test)
Basic average	0.077	0.107	0.213	0.271	0.102	0.095	0.092	0.170
Basic cumulative	0.086	0.117	0.116	0.278	0.154	0.099	0.102	0.162
Delta	0.072	0.069	0.332	0.254	0.087	0.117	0.071	0.198
Decay	0.073	0.082	0.086	0.152	0.104	0.094	0.078	0.109
Delta with ± α	0.081	0.069	0.106	0.192	0.081	0.095	0.075	0.119
Delta with decay	0.083	0.079	0.099	0.176	0.095	0.103	0.081	0.118
Delta with uncertainty	0.080	0.071	0.131	0.210	0.085	0.100	0.076	0.132
Decay with uncertainty	0.087	0.083	0.081	0.123	0.096	0.090	0.085	0.098

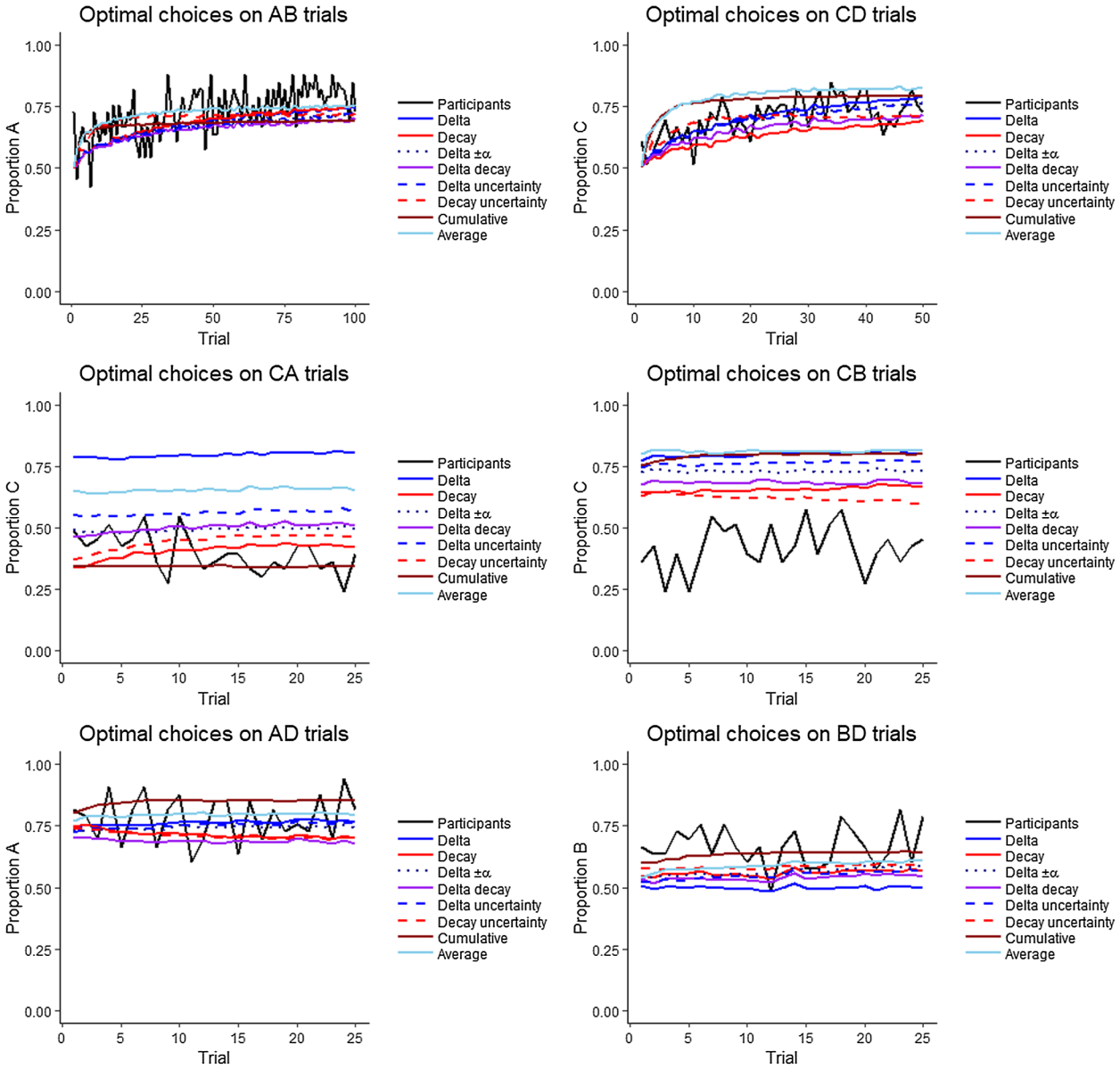

Fig. 9. Model choice predictions from *a priori* simulations against participants' actual choices shown by the solid black lines on (a) AB trials, (b) CD trials, and (c) CA trials.

capacity to recover the correct model. This involves simulating data from each model, and then fitting each model to the simulated data. Ideally, the model that generated the data will provide the best fit to the data; however, some models may be better able to mimic data produced by other models. We used both a data-uninformed cross-fitting procedure, which simulates data across the entire parameter space (reported below), as well as a data-informed cross-fitting procedure, which simulates data with parameters sampled from participants best-fitting parameter values (reported in the Supplemental Material). We focus on the data-uninformed version here, as it may be more

appropriate when comparing nested models (see Wagenmakers et al., 2004). In this study, we are comparing several models that are functionally similar. For example, we have several models based on average reward, and several based on cumulative reward. In this case, models may generate very similar patterns of data, particularly when using parameters that are fit to a particular pattern of data. It may therefore be difficult to distinguish models of the same type using this procedure, and as such, the results should be interpreted with caution.

In the data-uninformed cross-fitting procedure, we simulated 1000 data sets for each model for each experiment, and then fit each of the models to the simulated data and calculated BIC. Table 5 presents a confusion matrix that shows the proportion of simulated data sets that were best fit by each model across all three experiments, according to BIC. The only cases where a model did not provide the best fit of the data it generated were the Delta with decay model, which was better fit by the Decay model, and the Delta with ± α model, which was better fit by the Delta model. In both of these instances, the more complex model was better fit by a simpler model nested within it. This suggests that these models are generating similar data to their nested model, and that BIC is penalizing the additional complexity of the extended models.

When all possible models were considered, the Decay model provided the best fit for 83% of the data generated by the Decay model, while the Delta model provided the best fit for 33% of data generated by the Delta model. The lower proportion of data sets recovered by the Delta model is likely because there are more models based on average reward than cumulative reward, such that there is greater opportunity for the simulated Delta model data to be well fit by other models of the same type. To directly compare Delta and Decay models, we compared the fits of the Delta and Decay models, ignoring fits of other models, to data simulated by Delta and Decay models. The Delta model provided the best fit for 75% of the data generated by the Delta model, and the Decay model provided the best fit for 97% of the data generated by the Decay model. This suggests that neither model mimics the other model particularly well, and supports a main assertion of our paper that the Delta and Decay models predict fundamentally distinct behavior. However, the Decay model may be slightly more flexible than the Delta model when comparing these models directly with data-uninformed cross-fitting.

Because the Delta and Decay models are non-nested models of a different type, we also compared these models using the data-*informed* cross-fitting procedure where the parameters used to simulate the data are sampled from the best-fitting parameters from participants in this task. In this case, the Delta model provided the best fit for 86% of the data generated by the Delta model, and the Decay model provided the best fit for 85% of the data generated by the Decay model (see Supplemental Material). The models of primary interest therefore show similar levels of model recovery when using the best-fitting parameters. These proportions of data sets best fit by the data-generating model are similar to a previous paper from our lab that used the same cross-fitting method (Worthy, Otto, & Maddox, 2012).

Table 5 also shows the mean proportion of times each model provided the best fit to data that were generated by other models. The Delta and Decay models again do not appear to be overly flexible, providing the best fit to an average of only 8% and 9% of data sets produced by other models, respectively. We also calculated the mean proportion of data sets generated by models based on average reward that were best fit by each model, and the proportion of data sets generated by models based on cumulative reward that were best fit by each model. This allows us to further examine model flexibility by comparing how well models based on average reward could fit data generated by models based on cumulative reward, and vice versa. If the models are overly flexible, they will be able to provide a good fit to the data that are not produced by the same kind of model. Each model provided a better fit for data generated by the same type of model than a different type of model.

7. General discussion

This study tested the influence of reward frequency and average reward probability on choice in a reinforcement-learning task, comparing the predictions of learning rules based on average reward, and learning rules based on cumulative reward to test the underlying assumptions they make. We demonstrated that these models make divergent predictions about the value of alternative options when those options are presented with different frequency. Learning rules based on average reward, including the Delta rule, give greater value to options with a higher probability of reward, while learning rules based on cumulative reward, including the Decay rule, give greater value to options that have yielded more rewards overall. In our experiments, the critical test between these models was whether participants preferred option A, the more frequently rewarded option, or C, the option with the highest reward probability during the test phase. Most participants preferred option A on these trials, in support of the cumulative value models' predictions. The Decay model provided a better fit to the data than the Delta model, and was also able to reproduce this choice preference in *ex post* simulations. The results suggest that participants are more likely to base their decisions on how often each option had been rewarded, than on a learned estimate of the probability of receiving reward. In other words, the sensitivity to option frequency indicates that expected values of choice options are updated based on cumulative reward. This is in line with theories that suggest people do not learn reward probabilities directly, but instead store instances of reward associated with each option in memory and then translate these into choice probabilities that guide their behavior (Estes, 1976; Gonzalez & Dutt, 2011; Stewart, Chater, & Brown, 2006).

Table 5
Proportion of model-generated data sets best fit by each of the models.

	Fit model							
Simulated model	Average	Cumulative	Delta	Decay	Delta with decay	Delta with ± α	Delta with uncertainty	Decay with uncertainty
Average	**0.78**	0.05	0.02	0.02	0.01	0.06	0.05	0.01
Cumulative	0.00	**0.86**	0.00	0.05	0.00	0.05	0.01	0.02
Delta	0.18	0.03	**0.33**	0.05	0.26	0.10	0.05	0.01
Decay	0.05	0.08	0.00	**0.83**	0.01	0.00	0.00	0.02
Delta with decay	0.21	0.10	0.05	**0.36**	0.26	0.01	0.01	0.01
Delta with ± α	0.13	0.04	**0.43**	0.01	0.02	0.34	0.04	0.00
Delta with uncertainty	0.07	0.14	0.04	0.02	0.00	0.02	**0.65**	0.05
Decay with uncertainty	0.04	0.14	0.00	0.09	0.00	0.01	0.05	**0.67**
Fit to generative data	0.78	0.86	0.33	0.83	0.26	0.34	0.65	0.67
Fit to other data	0.10	0.08	0.08	0.09	0.04	0.04	0.03	0.02
Fit to average data	0.27	0.07	0.17	0.09	0.11	0.11	0.16	0.02
Fit to cumulative data	0.03	0.36	0.00	0.32	0.00	0.02	0.02	0.24

Note: The highest proportion of best-fit data sets for each simulated model is shown in bold.

The results are compelling as they reveal a clear influence of the amount of experience on choice option value, which is not predicted by one of the most popular learning rules. Delta-based models have also failed to predict other choice effects based on frequency differences. For example, the Rescorla-Wagner model cannot predict the choice preference for a rare outcome in the inverse base-rate effect (Markman, 1989). Delta-based learning is commonly used to model learning of action values from experience in diverse fields such as psychology (Otto & Love, 2010), computer science and neuroscience (e.g., McClure, Berns, & Montague, 2003). Beyond its ability to account for learning behavior, the Delta rule has continued to grow in popularity due to its ability to explain aspects of how dopaminergic brain regions encode prediction errors (Schultz & Dickinson, 2000; McClure et al., 2003; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Samanez-Larkin, Worthy, Mata, McClure, & Knutson, 2014). Given the prevalence of this model—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.

Adapting the Delta model to include additional parameters can allow it to better fit human behavior in our experiments. The addition of a decay parameter, which reduces the value of all options on every trial, allows the model to predict more A choices than the basic Delta model (see Collins & Frank, 2012). Including separate learning rates for positive and negative prediction errors also has a similar effect. The improvement in fit for models based on average reward with additional parameters means that the results cannot provide *exclusive* evidence for cumulative reward learning. Nevertheless, the results of these more complex models should be interpreted with caution. Additional parameters increase model flexibility; although the three-parameter models may provide improved fits to the data presented here, they are also more likely to be able to fit many other patterns of data than the two-parameter models (Roberts & Pashler, 2000). The Decay model can account for the data just as well, without the need for additional assumptions. In addition, these extended models tend not to predict a preference for A in ex post simulations, especially when the models are only fit to the training data. The models based on cumulative reward better predict choices on CA trials in these simulations.

An alternative explanation for the results is that the preference for A is simply an example of ambiguity aversion. Given C is experienced less often than A, its associated outcomes may be more uncertain. The uncertainty models used here reduce uncertainty the more frequently an option is chosen, and options with less uncertainty are valued higher when the uncertainty weight is negative. This mechanism differs from the decay model mechanism of reducing value as options are chosen less frequently. The Delta with uncertainty model provided a reasonably good fit to the data with a negative uncertainty weight. Within this data set, it is difficult to tease apart the uncertainty model's prediction that choosing an option more frequently reduces its uncertainty and the Decay model's prediction that value accumulates for options more frequently chosen. It is possible that both processes contribute, or that there are individual differences in whether people make choices based on accumulated value or reduced uncertainty. Uncertainty may play some role, for example, on CB trials; C is the clearly more valuable option, yet it is chosen less often than expected from the reward ratio. Future research will need to design tasks in order to more closely examine the influence of reward accumulation versus aversion to uncertainty. The Decay model and the Delta model with uncertainty could be used to identify tasks or paradigms where reward accumulation and aversion to uncertainty can be more directly dissociated.

Bayesian versions of the Delta rule model have recently been developed to account for uncertainty in addition to expected value by using a Kalman filter (Gershman, 2015). In the present work we simply used the variance of the beta distribution as our measure of uncertainty, but there are other potential metrics for representing uncertainty that could be tested in future work. It's possible that an improved metric for representing uncertainty might improve the Delta with uncertainty's models ability to account for the pattern of data we observed here. Additionally, some Bayesian reinforcement-learning models use uncertainty to replace the learning rate with the Kalman filter gain. This allows for more learning about less certain options. An interesting aim for future work might be to develop a Bayesian variant of the Decay model where something like the Kalman gain is used in the place of a decay parameter that is constant across trials.

Although our results support the Decay model there is still an extensive body of work that supports predictions from the Delta model (Rangel, Camerer, & Montague, 2008). A major finding is that prediction errors from the Delta model are correlated with activation of the ventral striatum (e.g. Hare et al., 2008; McClure et al., 2003). Additional work can be undertaken to identify whether neural activation in reinforcement-learning tasks 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. In light of the present findings, it is possible that many results previously found to support the Delta rule could also be accommodated by versions of a Decay rule. For example, we have found that mean-centered reward prediction errors from the Delta and Decay models are strongly correlated with each other. Thus, evidence for neural activity associated with reward prediction errors may not provide exclusive support for the Delta model, but instead reflects neural activity predicted by a variety of models. However, it is important to keep in mind that both models may have aspects that do not align with aspects of neurobiology or cognition (e.g. Steingroever, Wetzels, & Wagenmakers, 2014).

The Decay model may also need to be modified to be more generalizable. A more general version of the Decay model might cumulatively track the number of positive versus negative prediction errors, and would assume that participants make choices based on a recollection of positive versus negative outcomes associated with each option. Such a model would account for the frequency effects we observed here, and also account for the strong neural responses observed for prediction errors (e.g. Hare et al., 2008;). The Decay model presented here may also fail to account for reversal learning without additional assumptions added to the model (e.g. Boorman, Rajendran, O'Reilly, & Behrens, 2016), which should be addressed in future work.

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 (p.51). We found that people show a preference for options that have been more frequently rewarded than options with higher probability of reward. Thus, the current study provides evidence that people place greater value on cumulative instances of reward than on the probability of reward associated with different choice options. We further demonstrate that this is inconsistent with the basic Delta rule model, which tacitly assumes probability learning, and is instead more consistent with the predictions of the Decay model. A more complete theory to what has been outlined by the Decay model might posit that it is salience in memory, based on reward frequency, or any other factor, that drives choice. Outcomes that are more frequent are more memorable, but there could be other factors that influence the strength of associative memory between options available in the environment and positive outcomes associated with those options. Future research should also consider whether these results extend to variations in the amount of reward provided by each cue, rather than frequency. For example, whether there would be a preference for options that were associated with a higher amount of reward, but a lower probability of receiving that reward, when options are presented in equal frequency. Additionally, preferences for more frequently encountered options may also extend to other forms of learning, such as learning about causation. 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.

Open practices statement

The experiment, simulation and code are available on the Open Science Framework: https://osf.io/v57wf/.

Acknowledgements

This work was supported by grant AG043425 from the National Institute of Aging (NIA), United States to DAW. We thank research assistants Shannon Yap, Tuyet Linh Huynh, Sumedha Rao, Kirsten Downs, Ashton Wilson, Lilian Garza, Josh Philip, Mikayla Herman, Samantha Rumann, Kaila Powell, Kavyapriya Murali, Kinsey Blackburn, Shannon Pavloske, Marena De-Angelis, Catherine Lee, Melissa Hernandez, Tiffany Dobry, Xavier Jefferson, and lab manager Kaitlyn McCauley for assistance with the data collection.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.104042.

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Boorman, E. D., Rajendran, V. G., O'Reilly, J. X., & Behrens, T. E. (2016). Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. *Neuron, 89*, 1343–1354.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. *Journal of Personality and Social Psychology, 63*, 545–552.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Berlin: Springer.

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.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.

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

Collins, A. G. E., & Frank, M. J. (2012). How much reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35*(7), 1024–1035.

Curley, S. P., Yates, J. F., & Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes, 38*(2), 230–256.

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

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics, 75*, 643–669.

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.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLOS Computational Biology, 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*(4), 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*(22), 5623–5630.

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

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour, 1*, 0067.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General, 118*(4), 417–421.

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.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.). *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General, 145*(5), 548–558.

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.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences, 21*, 425–433.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7*, e1001048.

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.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347–353.

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.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

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). Adult age differences in frontostriatal representation of prediction error but not reward outcome. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 672–682.

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

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

St-Amand, D., Sheldon, S., & Otto, A. R. (2018). Modulating episodic memory alters risk preference during decision-making. *Journal of Cognitive Neuroscience, 30*(10), 1433–1441.

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. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1–26.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135–170.

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.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Meerhoff, F. (2018). Bayesian inference for psychology Part II: Example applications with JASP. *Psychonomic Bulletin and Review, 25*, 58–76.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50.

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

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

Witt, E. A., Donnellan, M. B., & Orlando, M. J. (2011). Timing and selection effects within a psychology subject pool: Personality and sex matter. *Personality and Individual Differences, 50*, 355–359.

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

Worthy, D. A., Maddox, W. T., & Markman, A. B. (2008). Ratio and difference comparisons of expected reward in decision-making tasks. *Memory & Cognition, 36*, 1460–1469.

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1640–1658.

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.