Journal of Mathematical Psychology 59 (2014) 41–49

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes

Darrell A. Worthy [a,*], W. Todd Maddox [b]

[a] Texas A&M University, United States
[b] The University of Texas at Austin, United States

HIGHLIGHTS

- Modifies a win-stay-lose-shift (WSLS) model using equations developed by W.K. Estes.
- Develops a dual-process WSLS–Reinforcement Learning (RL) model of decision-making.
- The WSLS–RL model accounts for behavior in three different decision-making tasks.
- Supports Estes' view of cognition consisting of multiple concurrent processes.

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
3 October 2013
Available online 8 November 2013

Keywords:
Decision-making
Dual-process
Mathematical modeling
Win-stay-lose-shift
Reinforcement learning

ABSTRACT

W.K. Estes often championed an approach to model development whereby an existing model was augmented by the addition of one or more free parameters to account for additional psychological mechanisms. Following this same approach we utilized Estes' (1950) own augmented learning equations to improve the plausibility of a win-stay-lose-shift (WSLS) model that we have used in much of our recent work. We also improved the plausibility of a basic reinforcement-learning (RL) model by augmenting its assumptions. Estes also championed models that assumed a comparison between multiple concurrent cognitive processes. In line with this, we develop a WSLS–RL model that assumes that people have tendencies to stay with the same option or switch to a different option following trials with relatively good ("win") or bad ("lose") outcomes, and that the tendencies to stay or shift are adjusted based on the relative expected value of each option. Comparisons of simulations of the WSLS–RL model with data from three different decision-making experiments suggest that the WSLS–RL provides a good account of decision-making behavior. Our results also support the assertion that human participants weigh both the overall valence of the previous trial's outcome and the relative value of each option during decision-making.

1. Introduction

The influence of W.K. Estes' work on the fields of Mathematical and Cognitive Psychology cannot be overstated. His pioneering work on verbal conditioning, which would later come to be known as probability learning, presaged work in reinforcement learning and reward-based decision-making that is extremely popular today. Central to Estes' work was the goal of explaining behavior in mathematical terms that could be formally modeled. He viewed the development and application of mathematical models of psychological phenomena as "a critical step in moving from descriptions of phenomena in ordinary language to representations in a theoretical plane" (Estes, 2002).

Another central theme in Estes' work was the notion of multiple concurrent processes in cognition (Estes, 1997, 2002; Estes & Da Polito, 1967; Maddox & Estes, 1996, 1997). He discussed this idea in much of his work on decision-making, recognition, and category-learning and made several attempts to formally model learning and memory processes by assuming that a comparison was made between the output of multiple concurrent cognitive processes, and the output of this comparison was what ultimately led to a response. The notion of multiple concurrent processes is a perennial theme in experimental psychology, and Estes was among those who championed this approach (Sloman, 1996; Smith & Decoster, 2000; Wason & Evans, 1975).

Much of our own recent work has centered on comparing fits of two different types of models to decision-making data: associative-based Reinforcement Learning (RL) models, and heuristic, or rule-based Win-Stay-Lose-Shift (WSLS) models (Cooper, Worthy, Gorlick, & Maddox, 2013; Worthy, Hawthorne, & Otto, 2013; Worthy & Maddox, 2012; Worthy, Otto, & Maddox, 2012). RL models have perhaps been the most popular models of decision-making

* Correspondence to: Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235, United States.
E-mail address: worthyda@tamu.edu (D.A. Worthy).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.10.001

over the past several decades and have been used to describe behavior in a number of different decision-making tasks (Erev & Roth, 1998; Frank, Seeberger, & O'Reilly, 2004; Sutton & Barto, 1998; Yechiam & Busemeyer, 2005). WSLS models have also been popular for quite some time, but have typically only been applied to data from binary choice experiments (Goodnow & Pettigrew, 1955; Medin, 1972; Novak & Sigmund, 1993; Otto, Taylor, & Markman, 2011; Steyvers, Lee, & Wagenmakers, 2009). Our recent work has demonstrated that WSLS models can often provide equally good or superior fits compared to RL models for data from a wide variety of decision-making tasks (Worthy, Hawthorne et al., 2013; Worthy & Maddox, 2012; Worthy et al., 2012).

In the current work we modify our WSLS model by utilizing equations first developed by Estes in his work modeling probability learning in the 1950s (Estes, 1957, 2002; Estes & Straughan, 1954). The modification significantly improves the fit of our WSLS model and allows the WSLS model to assume that tendencies to stay following a win or shift following a loss change over time. We also test an augmented version of a basic RL model. The basic RL model assumes that participants track the recency-weighted average rewards they receive when they select each option to determine each option's expected reward values. The recency-weighted averages, or expected reward values for each option, are then compared to determine the probability of selecting each option. The augmented version of the RL model, allows for the additional assumptions that participants may assign reward credit to options that were chosen in the recent past and that expected rewards for each option decay, or are "forgotten" as they are selected less often and. Recent work has demonstrated that adding these assumptions to the basic RL model can significantly improve the fit (Bogacz, McClure, Li, Cohen, & Montague, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Erev & Roth, 1998; Gureckis & Love, 2009; Howard-Jones, Bogacz, Yoo, Leonards, & Demetriou, 2010; Sutton & Barto, 1998).

We then combine the WSLS and RL models into a WSLS–RL Comparison model of decision-making inspired by Estes' later attempts to develop models of cognition that assumed multiple concurrent processes. The combined dual process model assumes: (a) that people have tendencies to stay with the same option or shift to a different option following trials with good (relative win) or bad (relative loss) outcomes (based on the WSLS model's assumptions), and (b) that the tendencies to stay or shift are adjusted based on the relative value of each option (based on the RL model's assumptions). Thus, people have tendencies to stay or switch on the next trial based on the overall outcome valence of that trial relative to the previous trial, and these tendencies are adjusted based on the value of the reward they expect to receive from each choice option. The model assumes that people are more likely to stay on a 'win' trial or shift on a 'loss' trial (WSLS), and they are more likely to stay-with or shift-to options with higher expected values than options with lower expected values (RL). The benefit of fitting a dual-process comparison model is that we can evaluate whether participants consider both the valence of the last outcome (WSLS) and relative value of each option (RL) to make decisions on each trial, rather than just the valence or the relative value.

Thus, the approach we take here is to augment two models that have been very successful in describing decision-making behavior by adding additional mechanisms. This is a common approach that was championed by Estes (1994):

> "*A standard, and very powerful, procedure that is available once we have a model that provides a good fit to a set of data is to augment the model by adding one additional mechanism or process of interest (often, but not necessarily, accomplished by adding one free parameter).... It is hard to overestimate the power of this technique for gaining evidence about mechanisms and processes that cannot be directly observed*".

In the following sections we first present the RL and WSLS models used as components in the WSLS–RL dual-process model, including the modification to our previous instantiation of the WSLS model based on Estes' early work in modeling probability learning (Estes, 1957; Estes & Straughan, 1954), and the modifications to the basic RL model to allow expected reward values to decay and for reward credit to be given to options chosen in the recent past. We then fit the dual-process WSLS–RL model to the data from three experiments and evaluate the degree to which weight is given to the valence of the prior outcome (WSLS) versus the relative value of each option (RL). We also simulate the model using best-fitting parameter values from participants in our experiments and compare the observed behavior of participants to that predicted by the WSLS–RL model. Our analysis included both a comparison of the proportion of times participants, and the model, select the most advantageous option over the course of the experiment, and how frequently participants, and the model, 'stay' by picking the same option that was selected on the previous trial, or 'shift' by picking a different option than the one chosen on the previous trial. This allows us to examine how the model accounts for both tendencies to select options with higher expected values and tendencies to stay or shift to different options depending on the outcome of the previous trial.

1.1. RL model

In decision-making situations involving choice, RL models assume that people develop Expected Values (EV) for each choice option that represent the reward (or punishment) they expect to receive following each choice. A probability for selecting each option a on trial t is typically given by a Softmax rule which provides an action selection probability for each option, a, based on its EV relative to the EVs of all j options (Sutton & Barto, 1998):

$$P(a_t) = \frac{e^{[\gamma \cdot EV_{a,t}]}}{\sum_{j=1}^{n} e^{[\gamma \cdot EV_{j,t}]}}. \quad (1)$$

Here γ is an exploitation parameter that determines the degree to which the option with the highest EV is chosen. As γ approaches infinity the highest valued option is chosen more often, and as γ approaches 0 all options are chosen equally often.

The basic RL model assumes that participants develop Expected Values (EVs) for each option that represent the rewards they expect to receive upon selecting each option. EVs for all options are initialized at $EV_{initial}$, a free parameter in the model at the beginning of the task, and updated only for the chosen option, i, according to the following updating rule:

$$EV_{i,t+1} = EV_{i,t} + \alpha \cdot [r(t) - EV_{i,t}]. \quad (2)$$

Learning is modulated by a learning rate, or recency, parameter (α), $0 \leq \alpha \leq 1$, that weighs the degree to which the model updates the EVs for each option based on the prediction error between the reward received ($r(t)$), and the current EV on trial t. As α approaches 1 greater weight is given to the most recent rewards in updating EVs, indicative of more active updating of EVs on each trial, and as α approaches 0 rewards are given less weight in updating EVs. When $\alpha = 0$ no learning takes place, and EVs are not updated throughout the experiment from their initial starting points. This model has been used in a number of previous studies to characterize choice behavior (e.g. Daw et al., 2006; Otto, Markman, Gureckis, & Love, 2010; Worthy, Maddox, & Markman, 2007; Yechiam & Busemeyer, 2005). The basic assumption behind RL models is that people probabilistically select options with higher EVs.

In the current work we augment the basic RL model in two ways. First, we allow the model to assume that eligibility traces for

recent actions determine the degree to which the credit for the reward received on each trial is given to each option. Eligibility traces assert that participants remember which options they have chosen in the recent past, and that some of the credit from the reward received on each trial goes to options chosen on previous trials, rather than all of the credit going to the option that was just chosen (Bogacz et al., 2007; Gureckis & Love, 2009; Sutton & Barto, 1998). Each time an option is chosen the eligibility trace for that option (i) is incremented according to:

$$\lambda_{i,t} = \lambda_{i,t-1} + 1. \tag{3}$$

*EV*s for all, j, options are then updated according to the following updating rule:

$$EV_{j,t+1} = EV_{j,t} + \alpha \cdot [r\,(t) - EV_{j,t}] \cdot \lambda_{j,t}. \tag{4}$$

On each trial, the eligibility trace, λ_j, for every option decays based on an eligibility trace decay parameter, ζ $(0 \leq \zeta \leq 1)$:

$$\lambda_{j,t+1} = \lambda_{j,t} \cdot \zeta. \tag{5}$$

Eligibility traces are meant to assert that participants remember which actions they have recently selected, and in this way recent actions can be credited if they lead to increases in reward on future trials. Higher decay parameter (ζ) values indicate less decay of memory traces for recent actions and more credit assignment to options that have been frequently selected in the recent past. The addition of eligibility traces to RL models has resulted in improved model performance in a variety of tasks (Bogacz et al., 2007; Gureckis & Love, 2009; Sutton & Barto, 1998).

The augmented RL model also allows the model to assume that EVs for each option are forgotten, or decay as they are selected less often (Howard-Jones et al., 2010). At the end of each trial the EVs for all j options are updated according to:

$$EV_{j,t+1} = EV_{j,t} \cdot \lambda + (1 - \lambda)\varepsilon \tag{6}$$

where λ $(0 \leq \lambda \leq 1)$ represents the rate of decay, with smaller values indicating faster decay, and ε represents the value to which EVs converge if an option is not chosen. The addition of the decay parameter to the model has provided a better fit to the data in much recent work (e.g. Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Erev & Roth, 1998; Howard-Jones et al., 2010; Worthy, Hawthorne et al., 2013). To summarize, on each trial the chosen option's eligibility trace is incremented to make the chosen option more "eligible" for learning (Eq. (3)), EVs for all options are updated based on the reward received (Eq. (4)), then both eligibility traces and EVs for all options decay (Eqs. (5)–(6)), and finally the RL model's probability of selecting each option is computed by comparing each options's EV (Eq. (1)).

1.2. WSLS model

An alternative strategy to the RL strategy of probabilistically selecting options expected to provide larger rewards is a WSLS strategy (Novak & Sigmund, 1993; Otto et al., 2011; Steyvers et al., 2009). WSLS is a rule-based strategy that has been shown to be commonly used in binary outcome choice tasks (e.g. Otto et al., 2011). Under this strategy, participants 'stay' by picking the same option on the next trial if they were rewarded, and 'shift' by picking the other option on the next trial if they were not rewarded.

This strategy can be modeled for data from binary outcome experiments like early work in probability learning (Estes & Straughan, 1954), but it can also be modeled for data from decision-making tasks where participants receive varying amounts of reward (or punishment) on each trial. In this more general form of the WSLS model participants "stay" by picking the same option on the next trial if the reward was equal to or larger than the reward received on the previous trial (a "win" trial), or "shift" by selecting the other option on the next trial if the reward received on the current trial was smaller than the reward received on the previous trial (a "lose" trial; Worthy & Maddox, 2012; Worthy et al., 2012).

The probabilities of staying following a "win" or shifting following a "loss" are free parameters in the model. In a two-alternative decision-making experiment the probability of staying with the same option, a, on the next t trial $(t+1)$ if the reward, r, received on the current trial is equal to or greater than the reward received on the previous trial is:

$$P\,(a_{t+1}|choice_t = a \text{ and } r\,(t) \geq r\,(t-1)) = P\,(stay|win)\,. \tag{7}$$

The probability of switching to another option following a win trial is 1-$P\,(stay|win)$.

The probability of shifting to the other option, b, on the next t trial $(t+1)$ if the reward, r, received on the current trial is less than the reward received on the previous trial is:

$$P\,(b_{t+1}|choice_t = a \text{ and } r\,(t) < r\,(t-1)) = P\,(shift|loss)\,. \tag{8}$$

The probability of staying with an option following a "loss" is 1-$P\,(shift|loss)$.

We have fit this model to experimental data in several of our recent studies, and it often provides a better fit than RL models (Worthy, Hawthorne et al., 2013; Worthy & Maddox, 2012; Worthy et al., 2012). However, one shortcoming of the model is that it is not a learning model because the best-fitting values of $P\,(stay|win)$ and $P\,(shift|loss)$ are estimated over all trials, and these values do not change throughout the experiment. It is reasonable to assume that the probability of staying on a "win" trial or shifting on a "loss" trial does not remain static over the course of the experiment.

In the early 1950s Estes encountered a similar situation when extending his statistical model for simple associative learning (Estes, 1950). In this model change in mean response probability on reinforced trials is given by:

$$p_{t+1} = p_t + \theta(1 - p_t). \tag{9}$$

Here the probability of a response increases on the next trial if a reward occurs on trial t, and θ performs a similar function that the learning rate (α) parameter performs in Eqs. (2) and (4). On unrewarded trials changes in mean response probability are given by:

$$p_{t+1} = (1 - \theta)p_t. \tag{10}$$

Here the probability of a response decreases on the next trial if a reward does not occur on trial t.

We utilized a modified version of Eq. (8) to modify $P\,(stay|win)$ and $P\,(shift|loss)$ on each trial based on whether the trial is a "win" or a "loss" trial. The modified WSLS model has six parameters: $P\,(stay|win)_{initial}$ and $P\,(shift|loss)_{initial}$, which represent the starting values of $P\,(stay|win)$ and $P\,(shift|loss)$, $P\,(stay|win)_{final}$ and $P\,(shift|loss)_{final}$, which represent the asymptotic ending values of $P\,(stay|win)$ and $P\,(shift|loss)$, and $\theta_{P(stay|win)}$ and $\theta_{P(lose|shift)}$ which determine how much $P\,(stay|win)$ and $P\,(shift|loss)$ change on each trial.

If $r\,(t) \geq r\,(t-1)$, then the trial is considered a "win" trial and the following equation that is of the same form as Eq. (8) is used to adjust $P\,(stay|win)$:

$$P\,(stay|win)_{t+1} = P\,(stay|win)_t + \theta_{P(stay|win)} \times (P\,(stay|win)_{final} - P\,(stay|win)_t). \tag{11}$$

If $r\,(t) < r\,(t-1)$, then the trial is considered a "loss" trial and

$$P\,(shift|loss)_{t+1} = P\,(shift|loss)_t + \theta_{P(lose|shift)} \times (P\,(shift|loss)_{final} - P\,(shift|loss)_t). \tag{12}$$

Modifying the WSLS model by adding Eqs. (8) and (9) allows the model to assume that participants' tendencies to stay or shift

on win and loss trials are modified throughout the experiment. This modification of the WSLS model allows the model to assume learning in that propensities to stay following a positive outcome or switch following a negative outcome are not required to remain static across all trials.

1.3. WSLS–RL model

RL and WSLS models can both capture behavior reasonably well in a variety of tasks. However, one possibility is that participants consider both the overall valence of the outcome on the previous trial (WSLS) and the relative value of each option (RL) to make decisions on each trial. The RL-based process provides information on the EV of each option relative to the EVs for all other options, while the WSLS-based process provides information on the participant's general propensity to stay with the same option or shift to a different option depending on whether the outcome was an improvement or a decline compared to the outcome on the previous trial. Modeling either process alone may not adequately account for human decision-making behavior. It is likely that human decision-making behavior involves a consideration of both the relative value of each option (RL) and the trend in rewards from trial to trial (WSLS).

The WSLS–RL model combines these two assumptions by assuming that the probability of selecting each option is affected by both the valence of the prior outcome and the relative value of each option. This assumption is accounted for by the WSLS–RL model by adding the parameter κ_{WSLS} which weighs the degree to which the WSLS model's output is utilized in determining the probability of selecting each j option:

$$P(j_t) = P(j_t)_{WSLS} \cdot \kappa_{WSLS} + P(j_t)_{RL} \cdot (1 - \kappa_{WSLS}). \quad (13)$$

This method of comparing the output from two separate models is one suggested by Estes in some of his later work (Estes, 2002; Maddox & Estes, 1996). In sum, the dual-process comparison WSLS–RL model has 13 free parameters, 6 for the RL-based process, 6 for the WSLS-based process, and κ_{WSLS} which weights the output of each process. The equations used for the WSLS–RL model are Eqs. (1), (3)–(8), and (11)–(12).

In the experiments presented below we fit the WSLS–RL Comparison model to data from three decision-making experiments that have quite different reward structures. We then examine the best fitting parameters from fits to subjects' data in our experiments. If people consider both the valence of recent outcomes (WSLS) and the relative value of each option then parameter estimates for κ_{WSLS} should be near 0.50. We also simulate the WSLS–RL model using best-fitting parameter estimates from participants in each experiment and compare the behavior of our experiments to that predicted by the model.

1.4. Overview of experiments

In Experiment 1 participants perform a binary outcome decision making task where they receive either three points or one point each time they select one of two options. One option provides the higher payoff (three points) 70% of the time and the other option provides the higher payoff only 30% of the time. In Experiment 2 participants perform a similar two choice task where they earn points on each trial and attempt to maximize the cumulative points earned. In this task one option provides an average payoff of 65 points on each trial, while the other option provides an average payoff of only 55 points on each trial. There is a standard deviation of 10 points around the average payoff for each option, and thus the task requires learning which option is better despite a high degree of noise in the rewards given by each option.

Experiments 1 and 2 have choice-history independent reward structures because the payoffs given on each trial are not influenced by the previous choices the participant has made. In Experiment 3 participants perform a choice-history dependent task where the payoffs are affected by the proportion of times participants have selected each option over the previous ten trials. One option, the *Increasing* option, causes future rewards for both options to increase, while the other option, the *Decreasing* option, causes future rewards for both options to decrease. The Increasing option is the optimal choice, but it always provides a smaller immediate reward compared to the decreasing option. Thus, the Decreasing option initially appears more rewarding despite being disadvantageous in the long run. Choice-history dependent tasks like these have recently become popular in examining how people avoid immediately rewarding options in favor of options that maximize long-term cumulative reward (Bogacz et al., 2007; Gureckis & Love, 2009; Otto et al., 2010; Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011).

2. Experiment 1

In Experiment 1 participants performed a two-choice binary outcome decision-making task where their goal was to maximize the cumulative points gained over the course of the experiment.

2.1. Method

2.1.1. Participants

Twenty young adults from the University of Texas at Austin participated in the experiment as partial fulfillment of a course requirement.

2.1.2. Materials and procedure

Participants performed the experiment on a PC using Matlab software with Psychtoolbox (Version 2.54). At the beginning of the task participants were told that they would select from one of two cards on each trial and that they would receive either one or three points upon each selection. The Advantageous deck gave three points with a probability of 0.7 and one point with a probability of 0.3, while the Disadvantageous deck gave three points with a probability of 0.3 and one point with a probability of 0.7.

On each trial participants were told to select one of the two options and were given as long as they wished to make a selection. Feedback was provided 500 ms after each response and lasted for 2000 ms before the next trial began. Participants performed 250 trials of the task. They were given a goal of trying to earn 600 points over the course of the task which is equivalent to earning the higher payoff (three points) on 70% of the trials.

2.2. Results

We fit the WSLS–RL model individually to each participant's data by maximizing log-likelihood. Table 1 lists the average best-fitting parameter values across all participants. For several parameters the distribution of best-fitting values across participants was markedly non-normal. The average best-fitting κ_{WSLS} was 0.48, which suggests that participants weighed the valence of the previous trial's outcome and the relative expected value of each option roughly equally. The distribution of best-fitting values for this parameter was more normal compared to other parameters from the model. The median of κ_{WSLS} parameter estimates was 0.47 and the 1st and 3rd quartiles were 0.39 and 0.60, respectively.

To examine the ability of the WSLS–RL model to account for participants' decision-making behavior we simulated the model using sets of best-fitting parameter values from participants in our experiment. We used the parameter values that best fit our participants' data for the simulated data sets. We generated 1000 data sets using parameter combinations that were sampled with replacement from the best-fitting parameter combinations for participants in our experiment. Thus, we randomly sampled a com-

Table 1
Average parameter values for the WSLS–RL model in Experiment 1.

	Mean parameter value	
α	0.40	(0.35, 0.11, 0.62)
γ	6.82	(3.82, 2.36, 9.57)
$EV_{initial}$	2.21	(0.91, 1.08, 3.00)
ζ	0.30	(0.31, 0.00, 0.47)
λ	0.85	(0.16, 0.70, 0.99)
ε	2.26	(0.85, 1.47, 3.00)
$P\,(stay\|win)_{initial}$	0.23	(0.32, 0.00, 0.34)
$P\,(shift\|loss)_{initial}$	0.60	(0.47, 0.00, 1.00)
$\theta_{P(stay\|win)}$	0.31	(0.42, 0.01, 0.63)
$\theta_{P(lose\|shift)}$	0.33	(0.44, 0.03, 0.88)
$P\,(stay\|win)_{final}$	0.92	(0.22, 0.97, 1.00)
$P\,(shift\|loss)_{final}$	0.37	(0.42, 0.00, 0.80)
κ_{WSLS}	0.48	(0.19, 0.39, 0.60)

Note: Numbers in parentheses represent standard deviations followed by the first and third quartiles.

bination parameters that provided the best fit to one participant's data and used those parameter values to perform one simulation of the task. We generated 1000 simulated data sets in this manner. This is the same approach that we have followed in recent work from our lab (Worthy, Hawthorne et al., 2013; Worthy et al., 2012; Worthy, Pang, & Byrne, 2013). This allowed us to compute the proportion of trials that the WSLS–RL model predicted participants would select the Advantageous option as well as the predicted number of trials that participants would 'stay' by picking the same option that they had picked on the previous trial.

Fig. 1(a) plots the average of the predicted and observed proportion of trials that participants selected the Advantageous deck in 25-trial blocks of the task. Error bars represent 95% confidence intervals. A repeated measures ANOVA showed a significant linear trend, $F(1, 19) = 20.07$, $p < 0.001$, partial $\eta^2 = 0.51$, and a significant quadratic trend $F(1, 19) = 7.52$, $p < 0.05$, partial $\eta^2 = 0.28$. Across all trials, participants selected the Advantageous deck on 72% of trials and earned an average of 548 points. Fig. 1(b) plots the average number of 'stay' trials in 25-trial blocks. A repeated measures ANOVA showed a significant linear trend, $F(1, 19) = 33.60$, $p < 0.001$, partial $\eta^2 = 0.64$, and a significant quadratic trend $F(1, 19) = 17.26$, $p < 0.01$, partial $\eta^2 = 0.48$. Overall, the WSLS–RL model provided a good account of behavior in the task. The predicted proportion of Advantageous deck selections were within the 95% confidence intervals for the observed choices from our participants in all ten 25-trial blocks in the task. The same was true for the predicted number of trials that participants selected the same option that they had selected on the previous trial.

2.3. Discussion

Behaviorally, and across all participants, there was evidence of "probability matching" where participants tended to select the Advantageous deck on about the same proportion of trials that deck gave the higher reward on (72% compared to 70% rate of higher payoff). Participants also tended to persevere, or select the same option that they had on the previous trial more as the experiment progressed. The WSLS–RL model's predicted proportion of Advantageous deck selections and stay trials corresponded closely to the proportion of times participants selected the Advantageous deck and persevered with the same option that they had selected on the previous trial over the course of the experiment.

Average estimated κ_{WSLS} parameter values were near 0.50 which suggests that decisions were made on the basis of both the expected value of each option and the overall valence of the previous trial's outcome. By combining the assumption that participants select options based on their relative expected values with the assumption that participants have a general tendency to stay or switch to a different option based on the outcome of the previous trial the WSLS–RL model was able to provide a good account of participants' behavior.

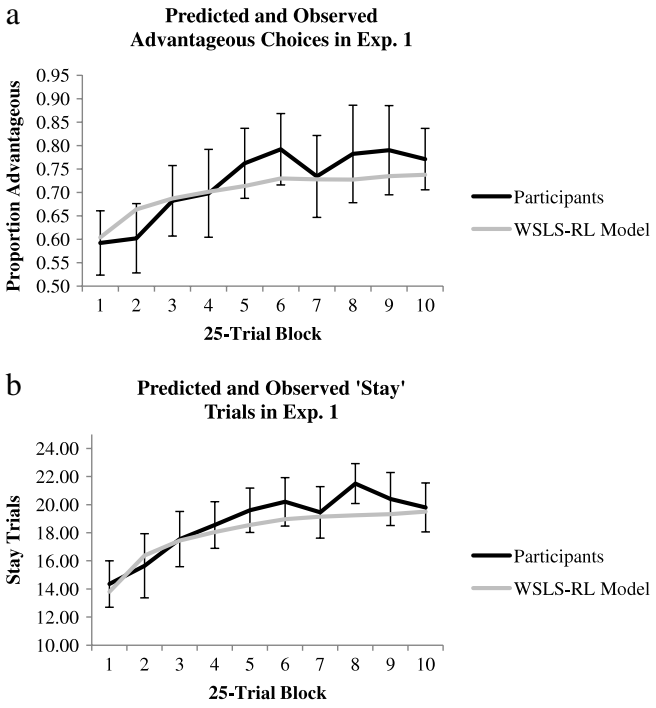

Fig. 1. (a) Predicted and observed proportion of trials that participants selected the advantageous option on in Experiment 1. (b) Number of predicted and observed proportion stay trials in Experiment 1. Stay trials were trials where the same option was selected that had been selected on the previous trial. Predictions were based on 1000 simulations of the WSLS–RL model. Error bars represent 95% confidence intervals.

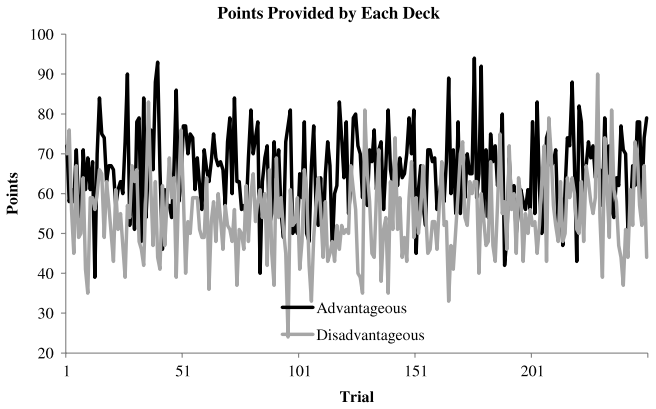

Fig. 2. Reward structure for Experiment 2.

3. Experiment 2

In Experiment 2 participants performed a decision-making experiment that shared many similarities with Experiment 1. However, the rewards in this task were continuously valued, rather than binary. Fig. 2 plots the rewards given by the Advantageous and Disadvantageous options on each trial. As stated above, mean payoffs of 65 and 55 points were given for the Advantageous and Disadvantageous decks, respectively. There was a standard deviation of 10 points around each deck's mean payoff.

3.1. Method

3.1.1. Participants

Twenty-three participants from the Texas A&M University community participated in the experiment in partial fulfillment of a course requirement.

Table 2
Average parameter values for the WSLS–RL model in Experiment 2.

	Mean parameter value	
α	0.32	(0.34, 0.03, 0.49)
γ	4.97	(3.95, 0.33, 8.34)
$EV_{initial}$	63.69	(22.55, 42.91, 89.52)
ζ	0.33	(0.33, 0.00, 0.63)
λ	0.84	(0.24, 0.82, 1.00)
ε	57.77	(24.88, 30 0.50, 84.87)
$P\ (stay\|win)_{initial}$	0.27	(0.37, 0.00, 0.46)
$P\ (shift\|loss)_{initial}$	0.68	(0.41, 0.36,1.00)
$\theta_{P(stay\|win)}$	0.30	(0.41, 0.03, 0.62)
$\theta_{P(lose\|shift)}$	0.28	(0.39, 0.03, 0.34)
$P\ (stay\|win)_{final}$	0.78	(0.33, 0.77,1.00)
$P\ (shift\|loss)_{final}$	0.28	(0.38, 0.00, 0.46)
κ_{WSLS}	0.46	(0.17, 0.33, 0.57)

Note: Numbers in parentheses represent standard deviations followed by the first and third quartiles.

3.1.2. Materials and procedure

Participants performed the experiment on a PC using Matlab software with Psychtoolbox (Version 2.54). At the beginning of the task participants were told that they would select from one of two cards on each trial and that they would receive between 1 and 100 points. On each trial participants were told to select one of the two options and were given as long as they wished to make a selection. Feedback was provided 500 ms after each response and lasted for 2000 ms before the next trial began. They were given a goal of collecting at least 16,000 points over the course of the experiment which could be reached by selecting the Advantageous deck on approximately 80% of the trials.

3.2. Results

Table 2 shows the average best-fitting parameter values for the WSLS–RL model. As in Experiment 1, many of the parameter estimates were non-normally distributed and the average best-fitting κ_{WSLS} was near 0.50 ($M = 0.46$), which suggests that participants weighed the valence of the previous trial's outcome and the relative expected value of each option roughly equally. The distribution of best-fitting values for this parameter was more normal compared to other parameters from the model. The median of κ_{WSLS} parameter estimates was 0.49 and the 1st and 3rd quartiles were 0.33 and 0.57, respectively.

We used the same bootstrapping method from Experiment 1 to simulate 1000 data sets for the WSLS–RL model using best-fitting parameter combinations from our participants. Fig. 3(a) shows the observed proportion of Advantageous deck selections over the course of the task along with predictions from the WSLS–RL model. A repeated measures ANOVA showed a significant quadratic trend $F(1, 22) = 28.12, p < 0.001$, partial $\eta^2 = 0.56$. Across all trials, participants selected the Advantageous deck on 73% of trials and earned an average of 15,762 points. Fig. 3(b) shows the predicted and observed number of stay trials over the course of the experiment. A repeated measures ANOVA showed a significant linear trend $F(1, 22) = 21.04, p < 0.001$, partial $\eta^2 = 0.49$, and a significant quadratic trend, $F(1, 22) = 11.54, p < 0.01$, partial $\eta^2 = 0.34$. As the task progressed learned to select the advantageous deck more frequently and they also switched options less frequently. The WSLS–RL model was able to account for both Advantageous deck selections and the number of trials that participants persevered by staying with the same option over the course of the task. Estimates for both of these measures were within 95% confidence intervals estimated from participants' data for all ten 25-trial blocks in the experiment.

3.3. Discussion

Participants selected the Advantageous deck more than the Disadvantageous deck, indicating a learned preference for the optimal deck in the task. Participants also tended to switch decks on successive trials less often as the task progressed. Parameter estimates suggested that participants' behavior was affected by both the relative expected value of each option and by the overall valence of the outcome on the previous trial. The WSLS–RL model was able to capture both the observed proportion of Advantageous deck selections and the number of trials that participants persevered by selecting the same option.

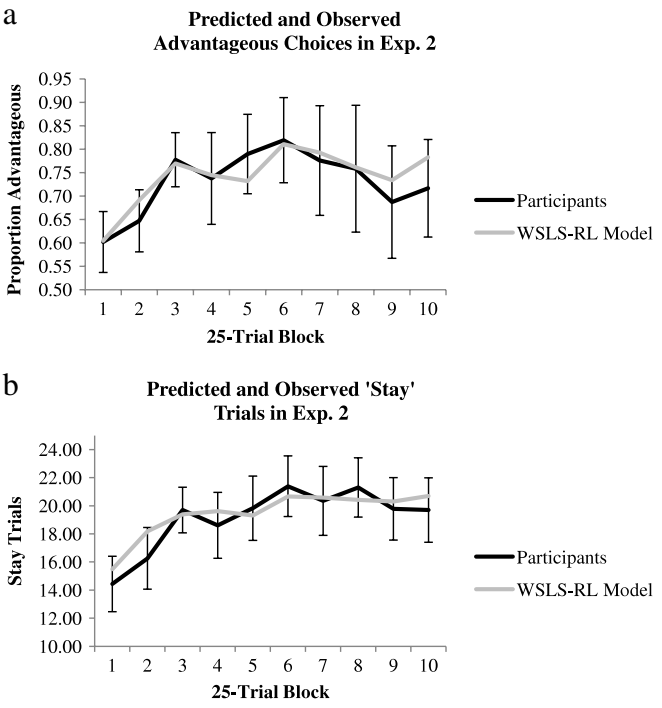

Fig. 3. (a) Predicted and observed proportion of trials that participants selected the advantageous option on in Experiment 2. (b) Number of predicted and observed proportion stay trials in Experiment 2. Stay trials were trials where the same option was selected that had been selected on the previous trial. Predictions were based on 1000 simulations of the WSLS–RL model. Error bars represent 95% confidence intervals.

4. Experiment 3

To provide a third test of the WSLS–RL model's ability to account for participants' decision-making behavior we had participants performing a dynamic, choice-history dependent decision-making task where the rewards given by each option depended on the recent choices participants had made. As in Experiments 1 and 2, participants performed a two-choice decision-making task where they were asked to pick from one of two decks of cards and maximize the cumulative points gained throughout the task.

The reward structure for the current task is shown in Fig. 4. The Increasing option provides a smaller immediate payoff on any given trial, but selecting this option causes participants to move to the right along the *x*-axis, and, as a result, earn higher payoffs regardless of which option they pick. In contrast, the Decreasing option always provides a larger immediate payoff, but selecting it causes participants to move to the left along the *x*-axis. Repeated selection of the Increasing option will lead to a reward of 80 points on each trial, while repeated selection of the Decreasing option will lead to a reward of only 40 points on each trial. Thus, the Increasing option is the advantageous choice for the task, while the Decreasing option is the disadvantageous choice. Good performance in the task requires forgoing the Decreasing option's larger immediate payoff in favor of the Increasing option's better long-term value. Thus, the task differs from the tasks in Experiments 1 and 2 where

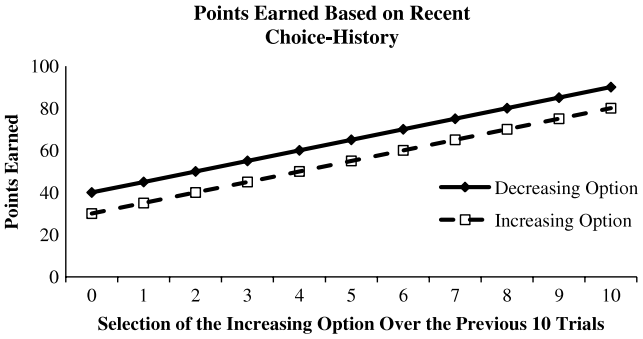

Fig. 4. Reward structure for the choice-history dependent task in Experiment 3. As indicated on the *x*-axis, the rewards provided by each option were determined by the number of times participants had selected the increasing option over the previous ten trials.

the payoffs were not affected by the choice-history of the participant, and the optimal choice maximized both immediate and cumulative reward.

4.1. Method

4.1.1. Participants

Twenty-three young adults from the Texas A&M University community participated in the experiment as partial fulfillment of a course credit.

4.1.2. Materials and procedure

Participants performed the experiment on PCs using Matlab software with Psychtoolbox (Version 2.54). At the beginning of the task participants were told that they would select from one of two cards on each trial and that they would receive between 1 and 100 points. On each trial participants were told to select one of the two options and were given as long as they wished to make a selection. Feedback was provided 500 ms after each response and lasted for 2000 ms before the next trial began. They were given a goal of collecting at least 18,000 points over the course of the experiment which could be reached by selecting the Increasing deck on approximately 80% of the trials.

4.2. Results

Table 3 shows the average best-fitting parameter values for the WSLS–RL model. As in Experiments 1 and 2, many of the parameter estimates were non-normally distributed and the average best-fitting κ_{WSLS} was near 0.50 ($M = 0.54$), which suggests that participants weighed the valence of the previous trial's outcome and the relative expected value of each option roughly equally. The distribution of best-fitting values for this parameter was more normal compared to other parameters from the model. The median of κ_{WSLS} parameter estimates was 0.56 and the 1st and 3rd quartiles were 0.36 and 0.71, respectively.

We used the same bootstrapping method used in Experiments 1 and 2 to simulate 1000 data sets for the WSLS–RL model using best-fitting parameter combinations from our participants. Fig. 5(a) shows the observed proportion of Advantageous deck selections over the course of the task along with predictions from the WSLS–RL model. A repeated measures ANOVA showed a significant linear trend $F(1, 22) = 13.47$, $p < 0.01$, partial $\eta^2 = 0.38$, and a significant quadratic trend $F(1, 22) = 8.45$, $p < 0.01$, partial $\eta^2 = 0.28$. Across all trials, participants selected the Advantageous deck on 52% of trials and earned an average of 15,174 points. Fig. 5(b) shows the predicted and observed number of stay trials over the course of the experiment. A repeated measures ANOVA showed a significant linear trend $F(1, 22) = 15.91$, $p < 0.01$, partial $\eta^2 = 0.42$, and a significant quadratic trend, $F(1, 22) = 6.85$, $p < 0.05$, partial $\eta^2 = 0.24$. As the task progressed learned to select the advantageous deck more frequently and they also switched options less frequently. The WSLS–RL model was able to account for both Advantageous deck selections and the number of trials that participants persevered by staying with the same option over the course of the task. Estimates for both of these measures were within 95% confidence intervals estimated from participants' data for every block in the experiment, except for the first block where the model over-predicted the proportion of Advantageous deck selections.

Table 3
Average parameter values for the WSLS–RL model in Experiment 3.

	Mean parameter value	
α	0.54	(0.40, 0.10, 0.97)
γ	4.98	(3.52, 0.92, 7.89)
$EV_{initial}$	59.81	(26.45, 30.47, 88.36)
ζ	0.58	(0.31, 0.43, 0.85)
λ	0.83	(0.25, 0.84, 0.97)
ε	49.62	(29.91, 32.46, 62.91)
$P\,(stay\|win)_{initial}$	0.38	(0.40, 0.00, 0.73)
$P\,(shift\|loss)_{initial}$	0.55	(0.42, 0.09, 1.00)
$\theta_{P(stay\|win)}$	0.20	(0.31, 0.02, 0.17)
$\theta_{P(lose\|shift)}$	0.33	(0.38, 0.04, 0.51)
$P\,(stay\|win)_{final}$	0.83	(0.33, 0.87, 1.00)
$P\,(shift\|loss)_{final}$	38	(0.35, 0.04, 0.66)
κ_{WSLS}	0.54	(0.22, 0.36, 0.71)

Note: Numbers in parentheses represent standard deviations followed by the first and third quartiles.

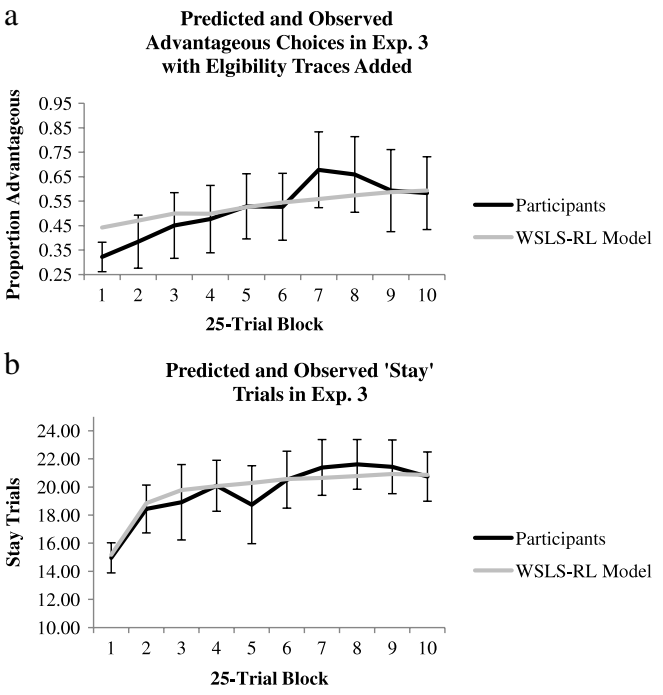

Fig. 5. (a) Predicted and observed proportion of trials that participants selected the advantageous option on in Experiment 3. (b) Number of predicted and observed proportion stay trials in Experiment 3. Stay trials were trials where the same option was selected that had been selected on the previous trial. Predictions were based on 1000 simulations of the WSLS–RL model. Error bars represent 95% confidence intervals.

4.3. Discussion

The dual process WSLS–RL model once again provided a good account for both the proportion of times participants selected the Advantageous deck and how frequently participants stayed with the same option over consecutive trials. However, the model over-predicted the proportion of Increasing option selections early in

the task. Initially, participants tended to select the disadvantageous Decreasing more frequently than the Increasing option. However, as the task progressed participants learned to select the Advantageous deck. In this respect, behavior in the choice-history dependent task used in Experiment 3 was slightly different than behavior in Experiments 1 and 2 where participants selected the Advantageous deck more often throughout the entire task. In Experiment 3 participants initially showed a bias toward the more immediately rewarding Decreasing option but learned to avoid it in favor of the Increasing option which provided more long-term cumulative reward.

On average, best-fitting κ_{WSLS} parameter values were again near 0.50 ($M = 0.54$) which suggests that the relative expected value of each option and the overall valence of the outcome on the previous trial affected decision-making behavior. One thing to note is that, on average, the decay parameter for eligibility traces (ζ) were higher in Experiment 3 ($M = 0.58$) than in Experiments 1 ($M = 0.30$) and 2 ($M = 0.33$). This is likely because eligibility traces are necessary for the model to value the Increasing option, which consistently provides smaller immediate rewards, over the Decreasing option (Gureckis & Love, 2009; Neth, Sims, & Gray, 2006). The addition of eligibility traces allow the model to assign reward credit to the Increasing option if it has been frequently chosen in the recent past and led to improvement in rewards for both options. Thus, in Experiment 3 eligibility traces slower decay rates for eligibility traces allow the model to assume that participants have greater memory for options chosen in the recent past which affect rewards on future trials. In this way the model learns to value the Increasing option over the Decreasing option.

5. General discussion

In three decision-making experiments that had qualitatively different reward structures the WSLS–RL dual process comparison model consistently provided a good account of participants' decision-making behavior. The model was able to account for both how often participants selected the Advantageous option and how often participants stayed with the same option over consecutive trials. This supports the assumption of the model that participants consider both the valence of the most recent outcome, as assumed by WSLS models, and the relative value of each option, as assumed by RL models, when making decisions. These are two psychologically plausible processes that mediate decision-making behavior and the dual process model makes the equally plausible assumption that participants consider both the valence of the previous outcome and the relative value of each option during decision-making. Thus, the dual process WSLS–RL Comparison model adds a missing component to each single-process model to better account for human behavior.

The weight given to the valence of the most recent outcome and the relative value of each option, which was estimated by the κ_{WSLS} parameter in the WSLS–RL model, was roughly equal in all three experiments. This suggests that participants may give roughly equal weight to the output of the valence of the reward on the last trial and the relative outcome of each option in a variety of decision-making contexts. Future work could investigate whether different experimental manipulations or individual differences among participants affect the degree to which participants' weigh the valence of the most recent outcome versus the relative value of each option.

The addition of the updating equations for the WSLS component of the model that were based on Estes' modification to his own learning model in the 1950s (Estes, 1957, 2002; Estes & Straughan, 1954) allowed the model to assume that tendencies to stay following a win or shift following a loss can change over time. The parameter estimates for $P(stay|win)_{initial}$, $P(stay|win)_{final}$, $P(shift|loss)_{initial}$, and $P(shift|loss)_{final}$, support the assumption that tendencies to stay or shift do indeed change over time. In each experiment the probability of staying on a win trial was greater at the end of the experiment than at the beginning, and the probability of shifting following a loss trial was smaller at the end of the experiment than at the beginning. Our analysis of the number of times participants stayed with the same option over the course of the task are in line with the parameter estimates from the WSLS–RL model in that, overall, participants tended to stay with the same option more often as the course progressed. This modification, directly inspired by Estes' work in the 1950s, also allowed the WSLS component of the model to assume that the stay and shift probabilities on each trial were adjusted based on feedback. Although the WSLS model can account for a wide range of data in decision-making experiments (Otto et al., 2011; Worthy & Maddox, 2012; Worthy et al., 2012; Worthy, Pang et al., 2013), it is very likely that "stay" and "shift" probabilities are dynamic and change throughout the course of the experiment.

Our approach of augmenting models that already provide a good account to experimental data by the addition of parameters that provide additional assumptions about behavior was an approach often taken and encouraged by Bill Estes in his own work (Estes, 1994, 2002). The result provides a powerful framework for research to test theories regarding what influences learning and decision-making behavior. Our approach has featured two prominent models of decision-making, but future work could test alternative augmentations along similar lines. There have been numerous different augmentations to models that employ the basic RL framework. Across a variety of domains researchers have fit RL models that assume eligibility traces for recent actions (Bogacz et al., 2007; Otto & Love, 2010; Sutton & Barto, 1998), attention to recent trends (Kovach et al., 2012) and perseverative autocorrelation (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), and this list is far from exhaustive. Additionally, there may be better ways to augments WSLS models. For example, the magnitudes of each "win" or "loss" could be considered when adjusted win-stay and lose-shift probabilities. We have focused on simple instantiations of the WSLS and RL-based processes in an attempt to isolate the components of each, and provide the clearest and most transparent test of the dual-process WSLS–RL model that we developed. Future work could test different augmentations of models that assume similar processes, or use a similar approach in entirely different domains. Such endeavors would be further testaments of the enduring legacy and footprint W.K. Estes left on the fields of Cognitive and Mathematical Psychology.

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.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*, 111–121.

Cooper, J. A., Worthy, D. A., Gorlick, M. A., & Maddox, W. T. (2013). Scaffolding across the lifespan in history-dependent decision-making. *Psychology and Aging*, *28*, 505–514.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (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. *The American Economic Review*, *88*, 848–881.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Estes, W. K. (1957). Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika*, *22*, 113–132.

Estes, W. K. (1994). *Classification and cognition*. Oxford: Oxford University Press.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*, 148–169.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*, 3–25.
Estes, W. K., & Da Polito, F. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology*, *75*, 18–26.
Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology*, *47*, 225–234.
Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: reinforcement learning in Parkinsonism. *Science*, *306*, 1940–1943.
Goodnow, J. J., & Pettigrew, T. F. (1955). Effect of prior patterns of experience upon strategies and learning sets. *Journal of Experimental Psychology*, *49*, 381–389.
Gureckis, T. M., & Love, B. C. (2009). Learning in noise: dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*, 180–193.
Howard-Jones, P. A., Bogacz, R., Yoo, J. H., Leonards, U., & Demetriou, S. (2010). The neural mechanisms of learning from competitors. *Neuroimage*, *53*, 790–799.
Kovach, C. K., Daw, N. D., Rudrauf, D., Tranel, D., O'Doherty, J., & Adolphs, R. (2012). Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *Journal of Neuroscience*, *32*, 8434–8442.
Maddox, W. T., & Estes, W. K. (1996). A dual-process model of category learning. In *Paper presented at the 31st annual meeting of the society for mathematical psychology*. University of North Carolina, Chapel Hill.
Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *3*, 539–559.
Medin, D. L. (1972). Role of reinforcement in discrimination learning set in monkeys. *Psychological Bulletin*, *77*, 305–318.
Neth, H., Sims, C. R., & Gray, W. D. (2006). Melioration dominates maximization: stable suboptimal performance despite global feedback. In R. Sun, & N. Miyake (Eds.), *Proceedings of the 28th annual meeting of the cognitive science society*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Novak, M., & Sigmund, K. (1993). A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*, 56–58.
Otto, A. R., & Love, B. C. (2010). You don't want to know what you're missing: When information about foregone rewards impedes dynamic decision making. *Judgment & Decision Making*, *5*(1), 1–10.
Otto, A. R., Markman, A. B., Gureckis, T. M., & Love, B. C. (2010). Regulatory fit and systematic exploration in a dynamic decision-making environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 797–804.
Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching. Evidence from a secondary task. *Cognition*, *118*, 274–279.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.
Smith, E. R., & Decoster, J. (2000). Dual-process model in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108–131.
Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning. *Cognition*, *3*, 141–154.
Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., & Maddox, W. T. (2011). With age comes wisdom: decision-making in younger and older adults. *Psychological Science*, *22*, 1375–1380.
Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: a comparison of win-stay-lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*, 364–371.
Worthy, D. A., & Maddox, W. T. (2012). Age-based differences in strategy-use in choice tasks. *Frontiers in Neuroscience*, *5*(145), 1–10.
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., 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*, Advance online publication.
Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation 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.