Published in final edited form as:
J Math Psychol. 2014 April 1; 59: 41–49. doi:10.1016/j.jmp.2013.10.001.

A Comparison Model of Reinforcement-Learning and Win-Stay-Lose-Shift Decision-Making Processes: A Tribute to W.K. Estes

Darrell A. Worthy[1] and **W. Todd Maddox**[2]

[1]*Texas A&M University*

[2]*The University of Texas at Austin*

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, and a comparison between the simple and more complex, augmented model determined whether the additions were justified. Following this same approach we utilized Estes' (1950) own augmented learning equations to improve the fit and plausibility of a win-stay-lose-shift (WSLS) model that we have used in much of our recent work. Estes also championed models that assumed a comparison between multiple concurrent cognitive processes. In line with this, we develop a WSLS-Reinforcement Learning (RL) model that assumes that the output of a WSLS process that provides a probability of staying or switching to a different option based on the last two decision outcomes is compared with the output of an RL process that determines a probability of selecting each option based on a comparison of the expected value of each option. Fits to data from three different decision-making experiments suggest that the augmentations to the WSLS and RL models lead to a better account of decision-making behavior. Our results also support the assertion that human participants weigh the output of WSLS and RL processes during decision-making.

Keywords

Decision-making; dual-process; mathematical modeling; win-stay-lose-shift; reinforcement learning

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).

Please address all correspondence to: Darrell A. Worthy Department of Psychology Texas A&M University 4235 TAMU College Station, TX 77843-4235 worthyda@tamu.edu.

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 (Wason & Evans, 1975; Sloman, 1996; Smith & Decoster, 2000).

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 (Worthy & Maddox, 2012; Worthy, Otto, & Maddox, 2012; Worthy, Hawthorne, & Otto, in press). RL models have perhaps been the most popular models of decision-making over the past several decades and have been used to describe behavior in a number of different decision-making tasks (Sutton & Barto, 1998; Erev & Roth, 1998; Frank, Seeberger, & O'Reilly, 2004; 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 (Novak & Sigmund, 1993; Medin, 1972; Goodnow & Pettigrew, 1955; Steyvers, Lee, & Wagenmakers, 2009; Otto, Taylor, & Markman, 2011). 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 et al., 2012; in press; Worthy & Maddox, 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 & Straughan, 1954; Estes, 1957; Estes, 2002). We refer to the old WSLS model as the $WSLS_{basic}$ model and the new WSLS model as the $WSLS_{learning}$ model. The modification significantly improves the fit of our WSLS model and allows the $WSLS_{learning}$ model to assume changes in response probability based on feedback, rather than static tendencies to stay or shift following relative win or relative loss trials. We then present a dual-process WSLS-RL 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 combined WSLS-RL dual-process model provides a superior fit to the data from three different decision-making tasks relative to either single-process model.

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) and then to compare, statistically when possible, the fits of the augmented and the "nested" original model, taking a significant difference in fit as evidence that the added mechanism or process pays its way and should be retained (Wickens, 1982; Estes, 1991). 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 & Straughan, 1953; Estes, 1957). We then fit the single process models and the dual-process WSLS-RL models to the data from three experiments to test whether augmentations to the single process models "pay their way" by significantly improving the fit to the data.

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* is typically given by a Softmax rule (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 0 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), $0 \leq a \leq 1$, 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.

1.2 WSLS Model

An alternative strategy to the RL strategy of probabilistically selecting options expected to lead to higher rewards is a WSLS strategy (Novak & Sigmund, 1993; Steyvers, Lee, & Wagenmakers, 2009; Otto, Taylor, & Markman, 2011). 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 "switch" 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, 1953), 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 on the next trial if the net gain received on the current trial is equal to or greater than zero is:

$$P\left(G_j\left(t\right)|choice_{t-1}{=}G_j\&r\left(t-1\right)\geq r\left(t-2\right)\right){=}P\left(stay|win\right) \quad (3)$$

In Equation 3 *r* represents the net payoff received on a given trial. The probability of switching to another option following a win trial is 1-*P*(*stay*|*win*).

The probability of shifting to the other option on the next trial if the reward received on the current trial is less than zero is:

$$P\left(G_j,\left(t\right)|choice_{t-1}{=}G_j\&r\left(t-1\right){<}r\left(t-2\right)\right){=}P\left(shift|loss\right) \quad (4)$$

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 et al., 2012; Worthy et al., in press; Worthy & Maddox, 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:

$$p_{t+1}=p_t+\theta\,(1-p_t) \quad (5)$$

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 Equation 2. On unrewarded trials changes in mean response probability are given by:

$$p_{t+1}=(1-\theta)\,p_t \quad (6)$$

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

We utilized these equations 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 $WSLS_{Learning}$ model has four parameters: $P(stay|win)_{initial}$ and $P(shift|loss)_{initial}$, which represent the starting 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 equations that are of the same form as Equations 5 and 6 are used to adjust *P*(*stay*|*win*) and*P*(*shift*|*loss*) respectively:

$$P(stay|win)_{t+1}=P(stay|win)_t+\theta_{P(stay|win)}\,(1-P(stay|win)_t) \quad (7)$$

$$P(shift|loss)_{t+1}=\left(1-\theta_{P(lose|shift)}\right)P(shift|loss)_t \quad (8)$$

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

$$P(shift|loss)_{t+1}=P(shift|loss)_t+\theta_{P(lose|shift)}\,(1-P(shift|loss)_t) \quad (9)$$

$$P(stay|win)_{t+1}=\left(1-\theta_{P(stay|win)}\right)P(stay|win)_t \quad (10)$$

Modifying the $WSLS_{Basic}$ model by adding Equations 7-10 allows the $WSLS_{Learning}$ model to assume that participants are more likely to continue to stay with the same option following successive "win" trials, and are more likely to shift to a different option following successive "loss" trials. This modification of the model also allows it to assume learning in

that propensities to stay following a positive outcome or switch following a negative outcome are not required to be static across all trials.

1.3 WSLS-RL Model

The RL and $WSLS_{Learning}$ models can both capture behavior reasonably well in a variety of tasks. However, one possibility is that participants compare output from RL-based and WSLS-based processes 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 $WSLS_{Learning}$ model provides information on the participant's general propensity to stay with the same option or switch 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 ($WSLS_{Learning}$).

The WSLS-RL model combines these two assumptions by assuming that participants weigh the output of RL and $WSLS_{Learning}$ processes in determining the probability of selecting each option. The parameter $W_{WSLS_{Learning}}$ weighs the degree to which the $WSLS_{Learning}$ model's output is utilized in determining the probability of selecting each option:

$$P(a_t) = P(a_t)_{WSLS_{Learning}} \cdot W_{WSLS_{Learning}} + P(a_t)_{RL} \cdot \left(1 - W_{WSLS_{Learning}}\right) \quad (11)$$

Here, $P(a_t)_{WSLS_{Learning}}$ is determined by Equations 3-4 and 7-10, and $P(a_t)_{RL}$ is determined by Equations 1 and 2.

1.4 Model Comparison Procedure

Table 1 summarizes the models we fit to the data in Experiments 1 – 3 and specifies the Equations used for each model. These include the RL and $WSLS_{Basic}$ models, along with the $WSLS_{Learning}$ model inspired by Estes' work, the dual process WSLS-RL model which weighs the output from the RL and $WSLS_{Learning}$ models, and a Baseline model that assumes a stochastic response pattern that does not depend on the reward given in previous trials. This model has one free parameter $p(a)$ that represents the probability of selecting option a on any given trial. The probability of selecting the other option on any trial is 1-$p(a)$ (Yechiam & Busemeyer, 2005).

In each experiment we fit each model to each participant's data individually based on the model's ability to predict the choice participants make on each trial by maximizing log-likelihood. We then compare Akaike weights for each model to determine the degree to which each of the five models listed in Table 1 provides the best fit to the data (Wagenmakers & Farrell, 2004). Akaike weights are derived from Akaike's Information Criterion (AIC) which is used to compare models with different numbers of free parameters (Akaike, 1974). AIC penalizes models with more free parameters. For each model, i, AIC is defined as:

$$AIC_i = -2logL_i + 2V_i \quad (12)$$

where L_i is the maximum likelihood for model i, and V_i is the number of free parameters in the model. Smaller AIC values indicate a better fit to the data. AIC values are then used to calculate Akaike weights to obtain a continuous measure of goodness-of-fit. A difference score is computed by subtracting the AIC of the best fitting model for each data set from the AIC of each model for the same data set:

$$\Delta_i(AIC) = AIC_i - minAIC \quad (13)$$

From the differences in AIC we then compute the relative likelihood, L, of each model, i, with the transform:

$$L(M_i|data) \propto exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\} \quad (14)$$

Finally, the relative model likelihoods are normalized by dividing the likelihood for each model by the sum of the likelihoods for all models. This yields Akaike weights:

$$w_i(AIC) = \frac{exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\}}{exp\left\{-\frac{1}{2}\Delta_k(AIC)\right\}} \quad (15)$$

These weights can be interpreted as the probability that the model is the best model given the data set and the set of candidate models (Wagenmakers & Farrell, 2004).

Nested Model Comparisons—Several of the models we include in our set are special cases nested within a more general model. The $WSLS_{Basic}$ model is a special case of the $WSLS_{Learning}$ model when $\theta_{P(stay|win)}$ and $\theta_{P(lose|shift)}$ both equal 0. The RL and $WSLS_{Learning}$ models are also nested within the WSLS-RL model when $W_{WSLSLearning}$ equals either 0 or 1. To directly evaluate the degree to which the augmentations of the RL and $WSLS_{Basic}$ models significantly improved the fit we also present G^2 comparisons between the $WSLS_{Basic}$ and $WSLS_{Learning}$ models, and between the RL and $WSLS_{Learning}$ models and the WSLS-RL model. The G^2 statistic for each model is given by:

$$G^2 = -2logL_i \quad (16)$$

The G^2 value of the general model, $G^2_{General}$, can never be larger than the G^2 value of the nested model, G^2_{Nested}. The statistic

$$G^2_{Nested-General} = G^2_{Nested} - G^2_{General} \quad (17)$$

can be computed, and this statistic follows a chi-square distribution with the degrees of freedom equal to the difference in the number of free parameters between the two models (Maddox & Ashby, 1993; Wickens, 1982).

Along with comparing Akaike weights in each experiment to evaluate which of the five models is most supported by the data we also compare the average $G^2_{Nested-General}$ values for each nested model comparison to the critical chi-square value based on the difference in the

number of parameters between each nested and general model. This approach allows us to determine if the dual process WSLS-RL model provides the best account to the data and the degree to which the augmentations to the RL and $WSLS_{Basic}$ models significantly improve fits to the data.

1.5 Overview of Experiments

The decision-making tasks we use assume qualitatively different reward structures. In Experiment 1 participants perform a binary outcome decision making task where they received 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 performed 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 (Gureckis & Love, 2009; Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011; Bogacz, McClure, Li, Cohen, & Montague, 2007; Otto, Markman, Gureckis, & Love, 2010).

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 & 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 .7 and one point with a probability of .3, while the Disadvantageous deck gave three points with a probability of .3 and one point with a probability of .7. 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. The best strategy in the task is to "maximize" by selecting the Advantageous deck on each trial, rather than to "probability match," or meliorate by selecting the Advantageous deck on 70% of the trials.

2.2 Results

On average, participants selected the Advantageous deck on 72% of trials and they earned 548 points across the task. We next examined the fits of each of the five models presented above to examine which model best described the decision-making behavior of our participants. The Akaike weights and average AIC values for each model are given in Table 2. Akaike weights were highest for the WSLS-RL model which indicates that it provided the best fit to the data.

Next, we performed G^2 comparisons between nested and general models. We first examined $G^2_{WSLS_{Basic}-WSLS_{Learning}}$ which indicates the degree to which the $WSLS_{Basic}$ model was improved by the additions of Equations 7-10. The $WSLS_{Learning}$ model has two additional parameters compared to the $WSLS_{Basic}$ model and the critical Chi-square value with 2 df and an α level of .05 is 5.99. The average $G^2_{WSLS_{Basic}-WSLS_{Learning}}$ across all participants was 11.56 (*SE*=3.69), which is well above the critical value, and the $WSLS_{Learning}$ provided a significantly better fit than the $WSLS_{Basic}$ model for 45% of the data sets from our participants.

For the comparison between the RL model and the WSLS-RL model the average $G^2_{RL-(WSLS-RL)}$ value across all participants was 29.31 (*SE*=7.88). This is much higher than the critical value of 11.07, with *df*=5 and α =.05, and the WSLS-RL model provided a significantly better fit for 70% of participants. The average $G^2_{WSLS_{Learning}-(WSLS-RL)}$ value across all participants was 17.67 (*SE*=2.51). The critical value was 7.82, with *df*=3 and α =. 05, and the WSLS-RL model provided a significantly better fit for 75% of participants.

We also examined the average best-fitting parameter values for the WSLS-RL, as well as the correlations between the best-fitting parameter values and performance. We used the proportion of times participants made the optimal choice, by selecting the Advantageous deck, over all trials as our measure of performance. The average best fitting parameter values and the correlations between each participant's best-fitting parameter values and performance are given in Table 3. The average $W_{WSLS_{Learning}}$ value was .50, indicating equal weight to the output of the $WSLS_{Learning}$ and RL models. $P(shift|loss)_{initial}$ and α(RL) parameter values were both significantly negatively correlated with performance, which suggests that higher probabilities of shifting following a loss trial and greater attention to

recent outcomes led to poorer performance. None of the other parameters were significantly associated with performance.

2.3 Discussion

Behaviorally, participants showed evidence of "probability matching" by selecting the Advantageous deck on about the same proportion of trials that that deck gave the higher reward on (72% compared to 70% rate of higher payoff). Our modeling results suggest that the dual process WSLS-RL model provided the best fit to the data overall. G^2 nested model comparisons suggest that the $WSLS_{Learning}$ model provided a significantly better fit than the $WSLS_{Basic}$ model, and that the WSLS-RL model provided a significantly better fit than either the RL model or the $WSLS_{Learning}$ model. Thus, the augmentations to the RL and $WSLS_{Basic}$ models 'paid their way' by significantly improving the fit.

The best-fitting parameter estimates suggest that participants equally weigh the output of the RL and $WSLS_{Learning}$ models. This is similar to what Otto and colleagues (2011) found in a comparable binary choice task in that both RL-based and WSLS-based processes can result in probability matching. Greater propensities to shift following a loss and greater attention to recent outcomes were both negatively associated with performance. This suggests that some switches away from the Advantageous deck may have been caused by receiving the lower payoff (one point) on the previous trial. Participants with higher estimated learning rate parameter values, α, from the RL model may have focused too heavily on recent outcomes which led them to undervalue the Advantageous deck after selecting did not lead to the higher payoff.

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. Figure 1 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.

3.1.2 Materials & 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. 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

On average, participants selected the Advantageous deck on 73% of trials and earned 15,761 points. Table 4 lists the Akaike weights and AIC values for each model. The WSLS-RL model provided the best fit to the data. Nested model comparisons showed that, on average, $G^2_{WSLS_{Basic}-WSLS_{Learning}}$ values were higher than the critical value of 5.99 (M=6.48, SE=1.09), with 48% of participants being significantly better fit by the $WSLS_{Learning}$ model than the $WSLS_{Basic}$ model.

Across all participants, the average $G^2_{RL-(WSLS-RL)}$ was higher than the critical value of 11.07 (*M*=42.85, *SE*=11.47), and the WSLS-RL model provided a significantly better fit for 74% of participants' data. The average $G^2_{WSLS_{Learning}-(WSLS-RL)}$ across all participants was also higher than the critical value of 7.82 (*M*=22.51, *SE*=3.64), and the WSLS-RL model provided a significantly better fit for 74% of participants' data.

Table 5 lists the average best-fitting parameter values for the WSLS-RL model as well as correlations with the number of trials participants selected the optimal, Advantageous deck. The average $W_{WSLS_{Learning}}$ value was .51, which suggests that participants equally weighed the output from the $WSLS_{Learning}$ and RL models. The $W_{WSLS_{Learning}}$ and γ(RL) parameters were also significantly negatively correlated with performance.

3.3 Discussion

Participants preferred the Advantageous deck more than the Disadvantageous deck, indicating a learned preference for the optimal deck in the task. The modeling results were similar to those from Experiment 1. There was evidence that the modifications to the $WSLS_{Basic}$ model based on Estes' probability updating equations (5 and 6) were justified, and the WSLS-RL model provided a better fit for a large majority of participants' data than either single-process model.

Participants' best-fitting $W_{WSLS_{Learning}}$ parameter values were negatively associated with performance which suggests that utilizing an RL strategy may lead to better performance in the task than utilizing a WSLS strategy. This is plausible because a WSLS strategy may have led to more switches away from the Advantageous deck since there were large changes in the rewards given from trial-to-trial. Participants' best-fitting γ(RL) parameters were also negatively correlated with performance. This relationship was unexpected because higher γ(RL) should lead to greater exploitation of the option with the higher expected value, which should have been the Advantageous deck.

To further investigate the unexpected relationship between estimated γ(RL) parameter values and the proportion of times participants selected the increasing option we examined the final expected values (EVs) from the RL model for each participant at the end of the task. We subtracted the EV of the Disadvantageous deck from the EV of the Advantageous deck which provided a measure of the degree to which participants valued the Advantageous deck over the Disadvantageous deck. We then examined the correlation between this measure of the relative value of the Advantageous deck with and the best-fitting γ(RL)

parameter values. There was a strong negative correlation between the relative values of the Advantageous deck and the best-fitting γ(RL) parameter values, $r=-.52$, $p<.001$. This suggests that the inverse relationship between γ(RL) values and performance may have been due to the model estimating higher γ(RL) when the value of the Advantageous deck, relative to the value of the Disadvantageous deck, was smaller. Thus, the negative correlation between γ(RL) values and performance was likely due to differences in the degree to which the model valued the Advantageous deck over the Disadvantageous deck rather than an inverse relationship between the degree to which participants exploited the option with the higher EV and performance. Higher γ(RL) values were estimated by the model to account for smaller differences in the EV of each option.

4. Experiment 3

To provide a third test of the WSLS-RL model's ability to better characterize choice behavior than either the RL or $WSLS_{Learning}$ single-process models we had participants perform 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 Figure 2. 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. The task involves 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 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 & 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. 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

On average, participants selected the Increasing option on 52% of trials and earned 15,174 points over the course of the Experiment. Table 6 lists the Akaike weights and AIC values for each model. Akaike weights were highest for the WSLS-RL model. Nested model comparisons showed that, on average, $G^2_{WSLS_{Basic}-WSLS_{Learning}}$ values were higher than the critical value of 5.99 (*M*=8.95, *SE*=3.05), with 52% of participants being significantly better fit by the $WSLS_{Learning}$ model than the $WSLS_{Basic}$ model.

Across all participants, the average $G^2_{RL-(WSLS-RL)}$ value was much higher than the critical value of 11.07 (*M*=121.55, *SE*=15.77), and the WSLS-RL model provided a significantly better fit for 96% of participants' data. The average $G^2_{WSLS_{Learning}-(WSLS-RL)}$ across all participants was also higher than the critical value of 7.82 (*M*=9.93, *SE*=1.84), and the WSLS-RL model provided a significantly better fit for 52% of participants' data.

Table 7 lists the average best-fitting parameter values for the WSLS-RL model and the correlations between best-fitting parameter values and proportion of trials in which participants selected the optimal, Increasing option. The average $W_{WSLS_{Learning}}$ parameter values were higher for participants in this Experiment (.68) than in Experiments 1 and 2 (.50 and .51), and there was also a positive association between estimated $W_{WSLS_{Learning}}$ parameter values and performance. There was also a positive relationship between estimated α(RL) parameter values and performance, and a significant negative correlation between estimated $\theta_{P(stay|win)}$ parameter values and performance.

4.3 Discussion

The dual process WSLS-RL model once again provided a better account for the data than either single process model. In this choice history-dependent decision-making task the $WSLS_{Learning}$ model's output was, on average, given greater weight than the RL model's output, and higher $W_{WSLS_{Learning}}$ parameter values were associated with better performance. This is consistent with previous work from our labs that found that a better fit of the $WSLS_{Basic}$ model, relative to fits of the RL model, was associated with better performance in the same task (Worthy et al., 2012). In this work we also found an association between higher learning rate parameter values (α) from the RL model and performance in this task. The negative association between $\theta_{P(stay|win)}$ parameter values and performance is likely due to lower performing participants being less likely to consistently stay with the Increasing option, which provides a series of consecutive 'wins' when it is repeatedly selected. Higher-performing participants were more likely to stay with the Increasing option, which was accounted for by the model with higher estimated $\theta_{P(stay|win)}$ parameter values.

5. General Discussion

In three decision-making experiments that had qualitatively different reward structures the WSLS-RL dual process comparison model consistently provided a better fit to the data than either single-process model. This supports the assumption of the model that participants compare the output of two different processes to arrive at a final probability of selecting

each option. The RL process computes choice probabilities for each option by comparing the EV of each, while the WSLS process computes a choice probability for each option based on whether the reward on the preceding trial was an improvement or a decline from the reward given two trials ago. Thus, the WSLS process provides information regarding the valence of the change in reward over the past two trials, while the RL process provides information regarding the recency-weighted expected value of each option. These are two psychologically plausible processes that mediate decision-making behavior and the dual process model assumes that the output of both processes is considered when making decisions.

The weight given to the output of the $WSLS_{Learning}$ and RL processes, which was estimated by the $W_{WSLS_{Learning}}$ parameter in the WSLS-RL model, was roughly equal in the two choice history-independent decision-making tasks used in Experiments 1 and 2. However, the WSLS model's output was given greater weight in the choice history-dependent decision-making task used in Experiment 3. This suggests that participants may give roughly equal weight to the output of the WSLS and RL processes when rewards are independent of their previous behavior, but that participants give greater weight to the output of the WSLS model when rewards rise or fall depending on which option they select. Future work could investigate whether different experimental manipulations or individual differences among participants affect the degree to which participants' weigh the output of each process.

The addition of the updating equations for the $WSLS_{Learning}$ model that were based on Estes' modification to his own learning model (Estes & Straughan, 1954; Estes, 1957; Estes, 2002) provided a significantly better fit to the data in all three experiments. This modification also allowed the $WSLS_{Learning}$ model to assume that the "stay" and "shift" probabilities on each trial were adjusted based on feedback, rather than static over the course of the experiment. Although the $WSLS_{Basic}$ model can account for a wide range of data in decision-making experiments (Worthy et al., in press; 2012; Worthy & Maddox, 2012; Otto et al., 2011), it is very likely that "stay" and "shift" probabilities are dynamic and change throughout the course of the experiment, which is supported by the better fits of the $WSLS_{Learning}$ model.

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, 2002; 1994). The result provides a powerful framework for research to test theories regarding what influences 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 (Sutton & Barto, 1998; Otto & Love, 2010; Bogacz et al., 2007), decay updating equations rather than the delta updating rule used in Equation 2 (Erev & Roth, 1998; Yechiam et al., 2005; Ahn, Busemeyer, Wagenmakers, & Stout, 2008), and perseverative autocorrelation (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), and this list is far from exhaustive. We have focused on simple instantiations of the WSLS and RL 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 WY, Busemeyer JR, Wagenmakers EJ, Stout JC. Comparison of decision learning models using the generalization criterion method. Cognitive Science. 2008; 32:1376–1402. [PubMed: 21585458]

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

Bogacz R, McClure SM, Li J, Cohen JD, Montague PR. Short-term memory traces for action bias in human reinforcement learning. Brain Research. 2007; 1153:111–121. [PubMed: 17459346]

Daw. ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69:1204–1215. [PubMed: 21435563]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441:876–879. [PubMed: 16778890]

Erev I, Roth AE. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. The American Economic Review. 1998; 88:848–881.

Estes WK. Toward a statistical theory of learning. Psychological Review. 1950; 57:94–107.

Estes WK. Theory of learning with constant, variable, or contingent probabilities of reinforcement. Psychometrika. 1957; 22:113–132.

Estes, WK. Statistical models in behavioral research. Erlbaum; Hillsdale, NJ: 1991.

Estes, WK. Classification and cognition. Oxford University Press; Oxford: 1994.

Estes WK. Processes of memory loss, recovery, and distortion. Psychological Review. 1997; 104:148–169. [PubMed: 9009883]

Estes WK. Traps in the route to models of memory and decision. Psychonomic Bulletin & Review. 2002; 9:3–25. [PubMed: 12026952]

Estes WK, Da Polito F. Independent variation of information storage and retrieval processes in paired-associate learning. Journal of Experimental Psychology. 1967; 75:18–26. [PubMed: 6065833]

Estes WK, Straughan JH. Analysis of a verbal conditioning situation in terms of statistical learning theory. Journal of Experimental Psychology. 1954; 47:225–234. [PubMed: 13152299]

Frank MJ, Seeberger LC, O'Reilly RC. By carrot or by stick: Reinforcement learning in Parkinsonism. Science. 2004; 306:1940–1943. [PubMed: 15528409]

Goodnow JJ, Pettigrew TF. Effect of prior patterns of experience upon strategies and learning sets. Journal of Experimental Psychology. 1955; 49:381–389. [PubMed: 14381587]

Gureckis TM, Love BC. Learning in noise: Dynamic decision-making in a variable environment. Journal of Mathematical Psychology. 2009; 53:180–193. [PubMed: 20161328]

Maddox WT, Ashby FG. Comparing decision bound and exemplar models of categorization. Perception and Psychophysics. 1993; 53:49–70. [PubMed: 8433906]

Maddox, WT.; Estes, WK. A dual-process model of category learning; Paper presented at the 31st Annual Meeting of the Society for Mathematical Psychology; Chapel Hill. Aug. 1996 University of North Carolina

Maddox WT, Estes WK. Direct and indirect stimulus-frequency effects in recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1997; 3:539–559.

Medin DL. Role of reinforcement in discrimination learning set in monkeys. Psychological Bulletin. 1972; 77:305–318.

Montague PR, Berns GS. Neural economics and the biological substrates of valuation. Neuron. 2002; 36:265–284. [PubMed: 12383781]

Novak M, Sigmund K. A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature. 1993; 364:56–58. [PubMed: 8316296]

Otto AR, Markman AB, Gureckis TM, Love BC. Regulatory fit and systematic exploration in a dynamic decision-making environment. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2010; 36:797–804.

Sloman SA. The empirical case for two systems of reasoning. Psychological Bulletin. 1996; 119:3–22.

Smith ER, Decoster J. Dual-process model in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and Social Psychology Review. 2000; 4:108–131.

Steyvers M, Lee MD, Wagenmakers EJ. A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology. 2009; 53:168–179.

Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. Psychonomic Bulletin & Review. 2004; 11:192–196. [PubMed: 15117008]

Wason PC, Evans J. St. B.T. Dual processes in reasoning. Cognition. 1975; 3:141–154.

Wickens, TD. Models for behavior. Freeman; San Francisco: 1982.

Worthy DA, Gorlick MA, Pacheco JL, Schnyer DM, Maddox WT. With Age Comes Wisdom: Decision-Making in Younger and Older Adults. Psychological Science. 2011; 22:1375–1380. [PubMed: 21960248]

Worthy DA, Hawthorne MJ, Otto AR. Heterogeneity of strategy use in the Iowa Gambling task: A comparison of win-stay-lose-shift and reinforcement learning models. Psychonomic Bulletin & Review. in press.

Worthy DA, Maddox WT. Age-based differences in strategy-use in choice tasks. Frontiers in Neuroscience. 2012; 5(145):1–10.

Worthy DA, Maddox WT, Markman AB. Regulatory fit effects in a choice task. Psychonomic Bulletin & Review. 2007; 14:1125–1132. [PubMed: 18229485]

Worthy DA, Otto AR, Maddox WT. Working-memory load and temporal myopia in dynamic decision-making. Journal of Experimental Psychology: Learning, Memory, and Cognition. Apr 30.2012 Advance online publication.

Yechiam E, Busemeyer JR. Comparison of basic assumptions embedded in learning models for experience based decision-making. Psychonomic Bulletin & Review. 2005; 12:387–402. [PubMed: 16235624]

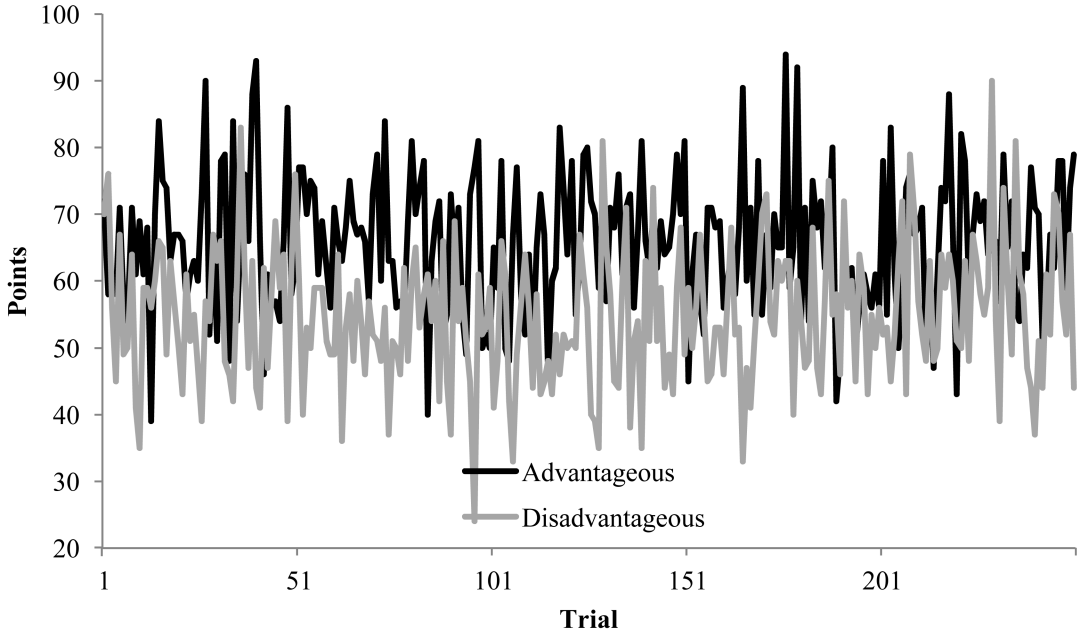

Figure 1.
Points Provided by Each Deck

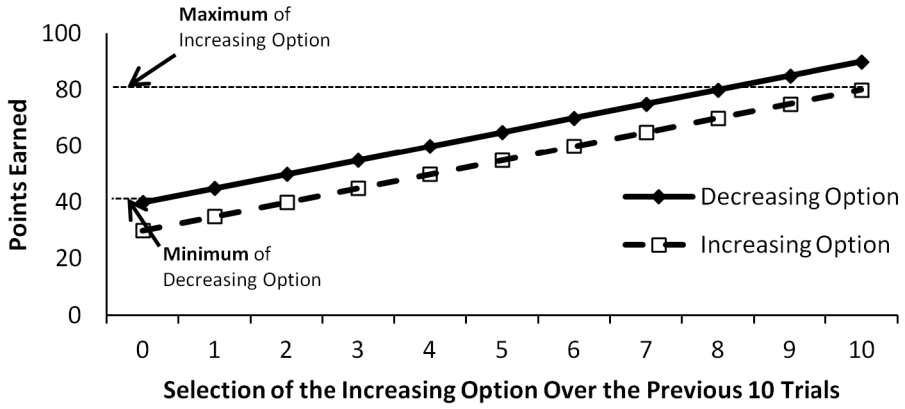

Figure 2.
Points Earned Based on Recent Choice-History

Table 1

Summary of Models Fit to Experimental Data

	Equations Used
RL Model	1-2
$WSLS_{Basic}$	3-4
$WSLS_{Learning}$	3-4, 7-10
WSLS-RL Model	1-4, 7-11
Baseline Model	

Table 2

Average Akaike Weights and AIC Values for Each Model from Fits to Experiment 1 Data

	Akaike Weights	Average AIC
RL Model	.24 (.08)	250.02 (11.47)
$WSLS_{Basic}$	.08 (.06)	249.93 (16.10)
$WSLS_{Learning}$	.08 (.05)	242.37 (15.36)
RL-WSLS Model	.48 (.10)	230.70 (15.24)
Baseline Model	.12 (.07)	280.02 (16.37)

Note: Numbers in parentheses represent standard errors of the mean.

Table 3

Average Parameter Values and Correlations with Proportion of Optimal Choices for the WSLS-RL Model

	Mean Parameter Value	Correlation with Performance
$P(stay\|win)_{initial}$	.75 (.08)	.03
$P(shift\|loss)_{initial}$	.30 (.07)	$-.52^{***}$
$\theta_{P\ (stay\|win)}$	.04 (.03)	.19
$\theta_{P\ (lose\|shift)}$	.40 (.11)	.13
α(RL)	.70 (.08)	$-.53^{***}$
γ(RL)	.95 (.05)	.09
$^{w}WSLS_{Learning}$	.50 (.05)	.18

Note: Numbers in parentheses represent standard errors of the mean.

* significant at p<.05;

** significant at p<.01,

*** significant at p<.001.

Table 4

Average Akaike Weights and AIC Values for Each Model from Fits to Experiment 2 Data

	Akaike Weights	Average AIC
RL Model	.22 (.07)	253.05 (16.09)
$WSLS_{Basic}$	.12 (.05)	239.19 (17.67)
$WSLS_{Learning}$	.14 (.06)	246.72 (17.49)
WSLS-RL Model	.50 (.09)	220.20 (17.38)
Baseline Model	.02 (.02)	263.79 (15.03)

Note: Numbers in parentheses represent standard errors of the mean.

Table 5

Average Parameter Values and Correlations with Proportion of Optimal Choices for the WSLS-RL Model

	Mean Parameter Value	Correlation with Performance
$P(stay\|win)_{initial}$	.90 (.04)	.32
$P(shift\|loss)_{initial}$	.28 (.07)	−.17
$\theta_{P\ (stay\|win)}$	.16 (.07)	.12
$\theta_{P\ (lose\|shift)}$	.39 (.10)	.−.11
α(RL)	.53 (.07)	−.17
γ(RL)	.29 (.07)	−.44*
W_{WSLS}	.51 (.06)	−.45*

Note: Numbers in parentheses represent standard errors of the mean.

** significant at p<.01,

*** significant at p<.001.

* significant at p<.05;

Table 6

Average Akaike Weights and AIC Values for Each Model from Fits to Experiment 3 Data

	Akaike Weights	Average AIC
RL Model	.04 (.04)	327.30 (9.26)
$WSLS_{Basic}$	.24 (.07)	224.62 (12.81)
$WSLS_{Learning}$	.24 (.06)	219.67 (13.20)
WSLS-RL Model	.44 (.08)	215.75 (13.93)
Baseline Model	.04 (.03)	288.90 (12.73)

Note: Numbers in parentheses represent standard errors of the mean.

Table 7

Average Parameter Values and Correlations with Proportion of Optimal Choices for the WSLS-RL Model

	Mean Parameter Value	Correlation with Performance
$P(stay\|win)_{initial}$	.81 (.05)	−.22
$P(shift\|loss)_{initial}$	.38 (.06)	−.15
$\theta_{P(stay\|win)}$	.19 (.06)	$-.60^{***}$
$\theta_{P(lose\|shift)}$	.48 (.10)	.05
α(RL)	.61 (.09)	$.48^{**}$
γ(RL)	.41 (.08)	−.07
W_{WSLS}	.68 (.06)	$.67^{***}$

Note: Numbers in parentheses represent standard errors of the mean.

* significant at p<.05;

** significant at p<.01,

*** significant at p<.001.