Running Head: PROCESSING REWARD

Processing Reward: How Humans and Learning Models Compare Different Expected Reward Values in Choice Tasks

Darrell A. Worthy, W. Todd Maddox, Arthur B. Markman

University of Texas, Austin

Word Count: 7,959

Please address all correspondence to:

Darrell A. Worthy
Department of Psychology
University of Texas
1 University Station, A8000
Austin, TX 78712
Phone: (512)475-8494
Fax: (512) 471-6175
worthyda@mail.utexas.edu

Abstract

Several models of choice compute the probability of selecting a given option by comparing the Expected Value (EV) of each option. However, there is a subtle difference between two common rules used to compute the action probability that is often ignored. Specifically, one common rule, the '*softmax*' rule compares the *distance* between the EVs, while another rule, the '*matching*' rule compares the *ratio* between the EVs. In this paper we test the assumptions of both rules by having human participants perform a choice task in which the reward values are shifted by an additive constant relative to a Control condition, so that the absolute distance between the EVs remains the same, or are multiplied by a constant relative to the Control condition, so that the ratio between the EVs remains that same. Results indicate that participants can more easily process the ratio than the absolute distance between the EVs. This finding has important implications for models of human choice behavior.

Keywords: Reward, Choice, Mathematical models, Gambling, Reinforcement Learning

Introduction

In choice tasks, such as the *n-armed bandit task*, a decision maker has to select an option in order to maximize benefit and minimize cost. For any given choice, this process often draws on past experience with the options. Several learning models have been proposed for this task that assume that decision makers compare the Expected Value (EV) of each option when determining the response to select on the next trial (Sutton and Barto, 1998; Busemeyer and Stout, 2002; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006, Worthy, Maddox, and Markman, 2007). These models use an action selection rule, derived from the biased choice rule (Luce, 1959, 1963; Shepard, 1957), for which the probability of selecting option x is equal to the EV for option x divided by the sum of the EVs for all possible options.

It is important to understand how these action-selection rules influence behavior because of their ubiquity in models of choice. Rules of this type have been incorporated into models of choice behavior in animals (often in the form of the 'matching law'; Herrnstein, 1961; Corrado, Sugrue, Seung and Newsome, 2005; Sugrue, Corrado, & Newsome, 2004; Lau & Glimcher, 2005), similarity-based models of category learning (e.g., Reed, 1972; Medin & Schaffer, 1978; Nosofsky, 1986; Maddox and Ashby, 1993), and many connectionist models of action selection, (Minsky & Papert, 1968, 1988; Rumelhart and McClelland, 1986, Kruschke, 1992). However, there is often a subtle difference in the implementation of these rules in computational models that is largely ignored in the literature. Specifically, whereas some models compare the *ratio* between the values representing each alternative, other models compare the *distance* between the values representing each alternative. Despite this difference, decision rules are typically incorporated into models without regard to how other choices of decision rule would have affected the behavior of the model.

In this paper, we compare these decision rules theoretically and empirically. We start by examining ratio-preserving and distance-preserving rules to understand their predictions for people's choice behavior. Then we test these models empirically in two choice studies that use a variant of the 2-armed bandit problem. We fit the obtained data with choice models and then discuss the implications of these results for models of choice behavior.

Ratio-preserving and Distance-preserving rules

Equation 1 depicts a ratio-preserving action selection model. The probability of selecting option *a* on trial *t* is determined by comparing the *ratio* between the EV of choice *a* with the EV of choice *b* on trial *t*:

$$P_{a,t} = \frac{EV_t(a)}{EV_t(a) + EV_t(b)} \tag{1}$$

In Appendix 1 we show algebraically that the probability of selecting option *a* in Equation 1 is dependent on the ratio between the EVs of options *a* and *b*, and thus is unaffected by scalar multiplication of the EVs. We refer to models of this sort as Matching models. Models of animal behavior have often used matching rules to compare the animal's response probabilities to the actual reward probabilities of each option (Herrnstein, 1961; Sugrue et al., 2004; Lau & Glimcher, 2005; Williams, 1988). Several popular models of category learning have also used matching rules to compute the probability of each categorization response [e.g. the Generalized Context Model

(Nosofsky, 1986), the deterministic Exemplar model (Maddox and Ashby, 1993), and the continuous-dimension version of the RULEX model (Nosofsky & Palmeri, 1998)].

Another action selection rule often used in computational models computes the probability of selecting option *a* on trial *t* by comparing the *difference* between the EV of choice *a* and the EV of choice *b* on trial *t.* An example of a model using this type of *softmax* action selection rule is:

$$P_{a,t} = \frac{e^{(EV_t(a))}}{e^{(EV_t(a))} + e^{(EV_t(b))}} \tag{2}$$

In Appendix 1 we show algebraically that the probability of selecting option *a* in Equation 2 is dependent on the distance between the EVs of options *a* and *b*, and thus is unaffected by scalar addition to the EVs. We refer to models of this sort as Softmax models. This type of rule is often employed in models of choice for tasks like the *n*-armed bandit problem (e.g., Sutton and Barto, 1998, Daw et al., 2005; Worthy et al., 2007; Busemeyer and Stout, 2002; Yechiam, Busemeyer, Stout & Bechara, 2005) and other domains including human foraging behavior (Roberts and Goldstone, 2006) and category learning (Kruschke, 1992; Love, Medin, and Gureckis, 2004).

The Matching model and the Softmax model make different assumptions about how decisions are made. For example, consider Decisions 1 and 2 with different EVs (represented in points) for options *a* and *b*:

Decision 1: $EV_t(a) = 6$; $EV_t(b) = 4$

Decision 2: $EV_t(a) = 60$; $EV_t(b) = 40$

Here the EVs in Decision 1 have been multiplied by 10 to create the EVs for Decision 2 so that the ratio between the EVs remains the same (0.6), but the distance between the

EVs is much greater in Decision 2 than in Decision 1 (i.e. 2 versus 20 for Decisions 1 and 2 respectively). Thus, the Matching model yields the same probability of selecting option *a* in both decisions (0.60), whereas the Softmax model yields different probabilities for selecting option *a* for Decisions 1 (0.88) and 2 (0.99).

In contrast, consider a third Decision:

$$\text{Decision 3: EV}_t(a) = 86;\ \text{EV}_t(a) = 84$$

In this case the EVs from Decision 1 have been increased by 80 points which maintains the same distance between the two EVs as in Decision 1, but changes the ratio between the two EVs from Decision 1. The Matching model now gives a probability of selecting option *a* of 0.51 for Decision 3 which is much different from the probability of 0.60 given by the Matching model for Decision 1. However, the Softmax model gives the same probability for Decisions 1 and 3 (0.88).

The different assumptions made by the models raise an important empirical question: Are people more likely to make decisions based on the ratio or the difference between the EVs representing each option? In this article we address this question by conducting two behavioral experiments in which participants receive rewards that are analogous to the rewards given in Decisions 1, 2, and 3 each time they select from one of two decks of cards while attempting to maximize the number of points that they earn in the task. In each experiment participants in the 'Control' condition perform a choice task in which they receive between 1 and 10 points for each draw. Participants in the Multiplied condition perform the same task except with reward values that have been multiplied by ten so that they receive between 10 and 100 points for each draw (in increments of 10). Participants in the Shifted condition perform the same task as the

Control condition except with reward values that have been shifted by 80 points so that they receive between 81 and 90 points on each trial.

The Matching and Softmax models make contrasting predictions regarding how well participants in each condition will exploit the option with the highest EV. The Softmax and Matching models make opposite theoretical predictions for participants receiving the shifted reward values between 81 and 90 points in which the distances between the EVs are the same as those for the Control condition, but the ratios between the EVs are much smaller than those for the Control condition. If the Matching model is more theoretically aligned with human behavior, people given shifted reward values between 81 and 90 points should perform worse than those in the Control condition, because the Matching model predicts less exploitation of the option with the highest EV, because the probabilities for selecting either deck will be closer to .50.. In contrast, if the Softmax model is more theoretically aligned with human behavior there should be no difference between the participants in the Shifted and Control conditions.

The Softmax and Matching models make opposite predictions for participants receiving reward values between 10 and 100 points on each trial in which the ratio between the EVs is the same as the Control condition, but the distance between the EVs are much greater. According to the Softmax model people receiving these multiplied values should actually perform better than those receiving reward values between 1 and 10 points on a task that requires exploitation of the option with the highest EV. This is because the model predicts greater exploitation of the option with the highest EV for participants receiving between 10 and 100 points on each trial due to greater absolute distances between the reward values. However, the Matching model predicts no difference between these two conditions because the ratios between the EVs for each option are the same.

In the next (second) section of this paper we present the results from two experiments that manipulate reward values in a choice task that either preserves (a) the difference between reward values (The Shifted condition) or (b) the ratio between reward values (The Multiplied condition). The analyses focus on a comparison of performance across conditions. The third section focuses on quantitative model-based analysis of the data from each experiment with expanded versions of the Softmax and Matching models presented above. Finally, we discuss the implications of this work for cognitive models of choice.

Experiment 1

In Experiment 1, participants were asked to draw from one of two decks of cards on each of 80 trials. The decks were constructed so that optimal responding required exploiting the deck that currently gave the highest reward. Deck B gave the highest reward over the first 50 trials and Deck A gave the highest reward over the final 30 trials. In the Control condition the values given for Deck A averaged 3 three points during the first 30 trials, four points over the next 20 trials, and seven points over the last 30 trials, while the values given for Deck B averaged eight points over the first 30 trials, six points over the next 20 trials, and three points over the last 30 trials. Thus participants needed to draw from Deck B for the first 50 trials and then to draw from Deck A for the final 30 trials in order to perform optimally on the task.

To compare the predictions of the two choice rules, we manipulated the reward values on the decks between participants. Participants in the Control condition received between 1 and 10 points for each draw. The decks for the Shifted condition were constructed by adding 80 points to each reward value in the Control condition, so that the values ranged between 81 and 90. The decks for the Multiplied condition were constructed by multiplying each reward value in the Control condition by ten, so that the values ranged between 10 and 90.

To maximize the number of points earned in the task participants had to distinguish the high value deck from the low value deck in order to exploit the option with the highest expected reward on each trial. The Matching model predicts worse performance for the Shifted condition relative to the Control and Multiplied conditions, while the Softmax model predicts superior performance for the Multiplied condition relative to the Control and Shifted conditions.

Method

Participants

Thirty University of Texas students participated in the experiment for course credit or monetary compensation ($6 base pay). Participants were told that their goal was to earn as many points as possible and that they could earn a $2 monetary bonus if they exceeded a pre-specified performance criterion.

Materials

Participants performed the experiment on a personal computer using Matlab software. Two decks appeared on the bottom half of either side of the screen. After each draw the selected card was overturned and placed above the selected deck. On the right side of the screen the phrase “Points required for the bonus” was shown followed by the points required in each condition. Below this the phrase “Your points” was shown, followed by the number of points each participant had earned.

Ten participants were assigned to the Control, Shifted, and Multiplied conditions. The Control condition received reward values between 1 and 10 points on each draw, the Shifted condition received reward values between 81 and 90 points on each draw, and the Multiplied condition received reward values between 10 and 100 points on each draw. For the Control condition deck “A” gave values that averaged three points during the first 30 trials, four points over the next 20 trials, and seven points over the last 30 trials, while deck “B” gave values that averaged eight points over the first 30 trials, six points over the next 20 trials, and three points over the last 30 trials. The specific deck values were determined by choosing a random point value with the relevant mean and a standard deviation of .88 with all values being rounded to the nearest whole number. The reward values given on each trial were identical for the Shifted and Multiplied conditions, except that 80 points was added to each reward given for the Shifted condition, and each reward was multiplied by ten points for the Multiplied condition. For the Control condition the maximum number of possible points that could be earned was 570 and the performance criterion was set at 550 points to ensure that a highly exploitative strategy was required in order to earn the monetary bonus order to achieve the bonus criterion. The performance criterion was adjusted accordingly for the Shifted condition (6,950 points) and the Multiplied condition (5,500 points).

Procedure

On each of 80 trials participants drew from one of two decks of cards. After each draw the card was overturned, the number of points received was shown, and added to the total number of points earned listed on the right side of the screen. The last card drawn from each deck remained visible until another card was drawn from that deck. After participants completed the experiment they were given the bonus if they earned it.

Results

To compare the point totals across conditions, points earned by participants in the Shifted condition were scaled by subtracting 80 points from each reward, and points earned by participants in the Multiplied condition were scaled by dividing each reward by 10. Figure 1a shows the average number of adjusted points earned in each condition. A one-way ANOVA revealed a main effect of condition, $F(2,27)=4.27, p<.05, \eta^2=.24$. Pairwise comparisons revealed significantly better performance in the Control (M=530.67) condition than in the Shifted (M=497.34) condition, $F(1,18)=4.85, p<.05, \eta^2=.21$. Likewise, participants in the Multiplied (M=533.62) condition earned significantly more points than participants in the Shifted condition, $F(1,18)=5.80, p<.05, \eta^2=.24$. However, there was no significant difference between the Control and Multiplied conditions.

We next analyzed the proportion of optimal choices made by each participant on each trial. Recall that Deck B gave higher rewards than Deck A on the first 50 trials, and Deck A gave higher rewards than Deck B on the last 30 trials. Thus an optimal choice was defined as selecting Deck B for the first 50 trials and Deck A for the final 30 trials. Figure 1b shows the average proportion of optimal choices made by participants in each condition. The pattern is the same as that of the total adjusted points earned data. A one-way ANOVA revealed a significant main effect of condition, $F(2,27)=4.41, p<.05, \eta^2=.24$. Pairwise comparisons revealed that a significantly larger proportion of optimal choices was made in the Control (M=.87) condition than in the Shifted (M=.76) condition, $F(1,18)=4.56$, $p<.05$, $\eta^2=.20$, and in the Multiplied (M=.87) condition than in the Shifted condition, $F(1,18)=5.07$, $p<.05$, $\eta^2=.22$. However, there was again no difference between the Control and Multiplied conditions.

Discussion

These data support the predictions made by the Matching model and not the Softmax model. When the reward values were shifted so that the distance between EVs remained constant, participants' performance was significantly worse than in the Control condition. In contrast, when the reward values were multiplied, so that the ratio of the EVs remained constant, participants' performance was comparable to the Control condition. These data suggest that human learners have difficulty exploiting the option with the highest EV in situations where the ratio between the EVs is low. Moreover, the differential performance between the Control and Shifted conditions cannot merely be due to the Shifted condition having to process larger reward values, because participants in the Multiplied condition had comparably high reward values to process on each trial. Thus the Matching model's prediction that participants would be less likely to exploit the option with the highest expected value as the ratio between the most and least valuable options decreased was confirmed.

We did not find strong support for the prediction of the Softmax model that participants in the Multiplied condition would be more likely than participants in the control condition to exploit the option with the highest EV on each trial, and thus earn more points. However, participants in the Multiplied condition did earn slightly more points than participants in the Control condition, and they also made slightly more optimal choices. To further test the predictions of each model we designed a second experiment in which the optimal deck would switch more periodically. Our motivation behind Experiment 2 was a) to replicate the findings of inferior performance for participants in the Shifted condition, and b) to design an experiment where the difference between the relatively good and bad values would be more obvious. By having the optimal deck switch more periodically we hoped to make more salient which deck currently gave the highest reward. This might allow participants in the Shifted condition a better opportunity to exploit the option with the highest EV.

Experiment 2

In Experiment 1, participants in the Shifted condition were unable to exploit the option with the highest EV as well as participants in the Control and Multiplied condition. This inferior performance was probably due to a lessened ability for participants in the Shifted condition to adequately distinguish between the relatively high and low EVs. We hoped that having the decks switch more periodically, and thereby exposing participants to the wider range of available reward values would make the differences between the relatively high and low reward values more obvious. This change might give participants in the Shifted condition a better chance at exploiting the option with the highest EV.

In this Experiment there is an optimal and a sub-optimal deck that switches every ten trials. Specifically the optimal deck for each ten-trial epoch averages seven points per card, and the sub-optimal averages three points per card. In this experiment we examine whether the inferior performance of participants in the Shifted group is a robust phenomenon, or if a manipulation of the task can cause participants in the Shifted condition to perform as well as the Control group. We also test the prediction of superior performance for participants in the Multiplied condition compared to participants in Control condition that is predicted by the Softmax model. Although we found no difference between the Control and Multiplied conditions in Experiment 1, it is possible that this design will allow participants in the Multiplied condition to better exploit the option with the highest EV than participants in the Control condition. This finding would confirm the prediction of the Softmax model that greater absolute differences between the high and low EVs leads to greater exploitation of the good deck.

Method

Participants

Thirty University of Texas at Austin students participated in the experiment for course credit or monetary compensation ($6 base pay). Participants were told that their goal was to earn as many points as possible in order to earn a $2 bonus.

Materials

Ten participants were placed in each of the Control, Shifted and Multiplied conditions. For participants in the Control condition, Deck A gave reward values that averaged seven points per draw (SD=0.88), and Deck B gave reward values that averaged three points per draw (SD=0.88) over the first ten trials of the experiment. After ten trials the rewards given by the two decks reversed so that during trials 11-20 Deck B averaged seven points per draw and Deck A averaged three points per draw. The reward values reversed in this manner every ten trials throughout the remainder of the experiment.

As in Experiment 1 participants in the Shifted condition received the same reward values as participants in the Control condition on each trial, except that 80 points was added to each reward value so that they received rewards ranging from 81-90 points on each trial. Similarly, participants in the Multiplied condition received the same rewards as those in the Control condition, except that each reward value was multiplied by ten points so that they received rewards ranging from 10-100 points in increments of ten. The maximum number of points a participant in the Control condition could earn on the task was 560. The bonus was set at 525 points which meant that participants had to be very vigilant in selecting the deck with the highest payoff. The bonus was adjusted accordingly for participants in the Shifted condition (bonus=6,925 points) and for participants in the Multiplied condition (bonus = 5,250 points).

All other materials were the same as those used in Experiment 1.

Procedure

The procedure was identical to the procedure for Experiment 1.

Results

As in Experiment 1, we analyzed the total number of adjusted points earned by each participant in the task using the same procedure outlined in Experiment 1. Figure 2a shows the average number of adjusted points earned in each condition. An ANOVA revealed significant differences between conditions, $F(2,27)=8.18$, $p<.01$, $\eta^2=.38$. Pairwise comparisons revealed that once again participants in the Shifted (M=450.9) condition earned significantly fewer adjusted points than participants in the Control condition (M=485.1), $F(1,18)=5.15$, $p<.05$, $\eta^2=.22$, and than participants in the Multiplied condition (M=504.2), $F(1,18)=12.59$, $p<.01$, $\eta^2=.41$. Interestingly, we also found that participants in the Multiplied condition earned significantly more adjusted total points than participants in the Control condition, $F(1,18)=4.46$, $p<.05$, $\eta^2=.20$.

We also analyzed the percentage of optimal choices in each condition. In this Experiment an optimal choice entailed selecting from the deck that gave the highest payoff. Figure 2b shows the average proportion of optimal choices in each condition. A one-way ANOVA revealed a significant effect of condition, $F(2,27)=8.37$, $p<.01$, $\eta^2=.38$. Pairwise comparisons showed that participants in the Shifted condition (M=.643) made significantly fewer optimal choices than participants in the Control condition (M=.764), $F(1,18)=5.83$, $p<.05$, $\eta^2=.25$, and than participants in the Multiplied condition (M=.818), $F(1,18)=12.23$, $p<.01$, $\eta^2=.40$. Participants in the Multiplied condition made marginally significantly more optimal choices than participants in the Control condition, $F(1,18)=3.93$, $p<.10$, $\eta^2=.18$.

Discussion

In Experiment 2 we once again found an inability to exploit the option with the highest EV for participants with Shifted reward values in which the ratio between the EVs changes but the distance remains the same. Participants in the Shifted condition earned fewer points and made a smaller proportion of optimal choices on a task that required exploitation of the option with the highest EV. These results confirm the prediction made by the Matching model that as the ratio between the option with the highest EV to the option with the lowest EV decreases, the probability of exploiting the option with the highest EV decreases. In contrast, this result contradicts the assumption drawn from the Softmax model that the absolute distance, and not the ratio, between the EV of each option governs responding.

The Softmax and Matching models also make contrasting predictions with respect to the Multiplied and Control conditions. In these two conditions, the ratio between the

EVs remains the same, but the absolute distance between the EVs changes. For this situation, the Softmax model predicts that participants in the Multiplied condition should be more likely to exploit the option with the highest EV, because the distance between the highest and lowest EV is much greater than in the Control condition. In contrast, the Matching model predicts no difference between these two conditions because the ratio between the EVs received in each condition is the same. In Experiment 1 we found no difference on the behavioral measures between participants in the Control and Multiplied conditions. However, in Experiment 2 we found that participants in the Multiplied condition earned more points than participants in the Control condition, and made a marginally larger proportion of optimal choices throughout the task.

These combined results suggest that shifting the scale of the reward values (as in the Shifted condition) may lead to rewards that are not as perceptually discriminable, while magnifying the scale of the reward values (as in the Multiplied condition) may lead to rewards that are more perceptually discriminable. Although we have so far found stronger evidence supporting the predictions of the Matching model than the predictions of the Softmax model it is important to note that the Softmax model correctly predicted the advantage found for participants in the Multiplied condition in Experiment 2.

In the next section we present model-based analyses of the data that include expanded versions of the Softmax and Matching models presented above in Equations 1 and 2. We expand the models by adding recency and exploitation parameters so that the degree to which each participant is exploiting the option with the highest EV can be compared across conditions.

Model-Based Analyses

To apply the Matching and Softmax models presented in Equations 1 and 2 to the data we first need a mechanism for updating the EV's on each trial. We used an incremental update rule (Sutton and Barto, 1998) for updating an average EV_k of the *k* past *(r)* rewards:

$$EV_{k+1} = EV_k + \alpha[r_{k+1} - EV_k] \tag{3}$$

where α, a recency parameter, varies from 0 to 1. When α=1 Equation 3 reduces to

$$\mathrm{EV}_{k+1} = \mathrm{r}_{k+1} \tag{4}$$

so that only the most recent rewards are used to estimate the value of a response option, and as $\alpha \to 0$ Equation 3 reduces to

$$\mathrm{EV}_{k+1} = \mathrm{EV}_{k} \tag{5}$$

so that all rewards are more equally weighted.

Both models also include an exploitation parameter (γ) that estimates the degree to which the option with the highest EV is exploited. Higher values of γ indicate greater exploitation of the option with the highest EV, while lower values of γ indicate greater exploration of alternative options with lower EVs. The action selection rule used by the Matching model is presented in Equation 6:

$$P_{a,t} = \frac{EV_t(a)^{\gamma}}{\sum_{b=1}^{n} EV_t(b)^{\gamma}} \tag{6}$$

The action selection rule for the Softmax model is presented in Equation 7:

$$P_{a,t} = \frac{e^{(\gamma EV_t(a))}}{\sum_{b=1}^{n} e^{(\gamma EV_t(b))}} \tag{7}$$

In both of these equations the degree to which the option with the highest EV is exploited is dependent on the exploitation parameter (γ). Parameter estimates and model fits are obtained based on the maximum log-likelihood (Wickens, 1982) for predicting the choice on the next trial.

Analysis of Experiment 1

To compare parameter estimates we adjusted the reward values earned in each condition as we did to compare the total points earned across conditions above. After adjusting the data from participants in the Shifted condition by subtracting 80 points from each reward received we fit each participant's data using the Matching model. Table 1 shows the average exploitation parameter values for the Matching Model in each condition. Although participants in the Shifted (M=3.38) condition had lower exploitation parameter values than participants in the Control (M=4.13) and Multiplied (4.42) conditions, a one-way ANOVA by condition was not statistically significant, F<1. Pairwise comparisons between participants in the Control and Shifted conditions and participants in the Multiplied and Shifted conditions were not statistically significant, F<1.5 for both comparisons.

The right side of Table 1 shows the average adjusted exploitation parameter values estimated by the Softmax model for participants in each condition. A one way ANOVA was statistically significant, $F(2,27)=4.38$, $p<.05$, $\eta^2=25$. Pairwise comparisons revealed a significant difference between estimated exploitation parameter values for participants in the Control (M=.84) and Shifted (M=.49) conditions, $F(1,18)=7.90$, $p<.05$, $\eta^2=.31$, and a significant difference between estimated exploitation parameter values for participants in the Multiplied (M=.90) and Shifted conditions, $F(1,18)=6.55$, $p<.05$,

η^2=.27. However, there was no difference between recovered exploitation parameter values for the Control and Matching conditions.[1]

We also examined which model provided the best overall fit to the data. Because the Matching and Softmax models have the same number of free parameters the maximum log-likelihood fits of each model can be compared directly. Figure 3 shows the proportion of participants in each condition best fit by the Softmax model. Interestingly, more participants in the Multiplied condition were best fit by the Softmax model while more participants in the Shifted condition were best fit by the Matching model, although binomial tests were not significant. This suggests that participants in the Shifted condition are indeed paying more attention to the ratio between the EVs, and participants in the Multiplied condition are paying greater attention to the distance between the EVs.

Discussion

Fits of the Softmax model indicate that the superior performance for participants in the Control and Multiplied conditions is due to a greater ability to exploit the option with the highest EV. Participants in these two conditions had significantly larger exploitation parameter values than participants in the Shifted condition leading to a larger proportion of optimal choices and more points earned.

However, significant exploitation parameter differences emerged for the Softmax model and not for the Matching model. This may be due to a greater variance in recovered exploitation parameter values from fits using the Matching model than from fits using the Softmax model. This points to a potential failure of the Matching model to

[1] For both models we failed to find any differences between the estimated recency parameters (α from Equation 4 above).

differentiate more exploratory from less exploratory strategies, and may suggest that the Softmax model is more appropriate for analyzing behavior in n-armed bandit choice tasks.

Analysis of Experiment 2

We fit the Matching and Softmax models to adjusted data from Experiment 2. In this task the option with the highest reward payoff changed a total of seven times, or after every ten trials, so optimum performance required exploitation of the option with the highest EV based on the most recent rewards given by each deck. For this reason we expected to find very high recency parameter estimates indicating greater weight on recent rewards. This is exactly what we found. The lowest average recency parameter value across all conditions and both models was 0.95 (for participants in the Multiplied condition fit best by the Softmax model). All other average recency parameter estimates were between .95 and 1, indicating a reliance on only the most recent rewards when determining the EV of each option across all conditions.

The left column of Table 2 shows the average exploitation parameter values from the Matching model. Although participants in the Multiplied condition had slightly higher exploitation parameter values than participants in the other two conditions, a one-way ANOVA revealed no significant difference among the three conditions, $F<1$. The right column of Table 2 presents the average normalized exploitation parameter values from the Softmax model. Once again there are no reliable differences between the average exploitation parameter values across conditions, $F<1$.

Both models appear unable to account for the behavioral differences between participants in the three conditions. This difficulty may be due to the fact that the models only update the chosen option on each trial, and the EV for the unchosen option remains unchanged even when the reward given by the chosen option is well below its EV. It is plausible that the unchosen option appears more appealing to participants when the chosen option is less rewarding than expected, and the unchosen option appears less appealing when the chosen option is more rewarding than expected. This might especially be true in the current experiment when the option with the highest payoff changes so frequently.

To account for an updating of the unchosen option we developed an Equal Updating Softmax Model[2] that incorporated an equation that mirrored Equation 3 shown above. Here the $EV_{u(k+1)}$ of the unchosen option is

$$EV_{u(k+1)} = EV_{u(k)} + \beta[EV_{c(k)} - r_{c(k+1)}] \tag{8}$$

where $EV_{u(k)}$ is the previous EV of the unchosen option, $EV_{c(k)}$ is the previous EV of the chosen option, $r_{c(k+1)}$ is the reward given by the chosen option, and β is a recency parameter for the unchosen option. In this equation if the reward given by the chosen option is greater than the EV of the chosen option then the EV of the unchosen option decreases, and if the reward given by the chosen option is less than the EV of the chosen option then the EV of the unchosen option increases. This model is similar in spirit to Erev and Roth's (1998) decay model, but here the EV of the unchosen option can either decrease (i.e. decay) or increase as a function of the reward given by the chosen option. This model allows for more flexible updating of the EV of each option, and so it may

[2] We amended only the Softmax and not the Matching model because we found parameter differences that accounted for the behavioral differences seen in the data from Experiment 1 using the Softmax model, but found no differences using the Matching model.

provide better fits to the data than the Softmax model for the current task where the option with the best payoff switches so frequently.

We first examined the average best-fitting recency parameter values for the unchosen option (β in Equation 8 above) for each condition. Figure 4 shows the average recency parameter value for updating the EV of the unchosen option across the three conditions. Although, participants in the Control and Multiplied conditions had, on average, higher estimated values of this parameter than participants in the Shifted condition, the difference was not significant, due to the large variance in each condition for this estimated parameter value.

Because the Softmax and Equal Updating Softmax models are nested, we used the G^2 criterion (Wickens, 1982; Maddox and Ashby, 1993), where $G^2 = -2\ln L$, to determine whether the more general model provided a significant improvement in fit for each participant's data. The difference between the G^2 values of the specific and general versions of the model has an asymptotic chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

For the ten participants in each condition the Equal Updating Softmax model fit significantly better than the Softmax model for five participants in the Control condition, four participants in the Shifted condition, and six participants in the Multiplied condition. Figure 5 shows the average G^2 fit values for participants in each condition. Although participants in the Control (M=55.02) and Multiplied (M=48.61) conditions have lower G^2 values than participants in the Shifted (M=60.86) condition, a one-way ANOVA was not significant, $F>1$.

To determine whether there was an advantage in the task for participants who were best fit by the Equal Updating Softmax model we next compared the proportion of optimal choices made by each participants across all three conditions based on whether they were best fit by the Softmax or Equal Updating Softmax model. Figure 6a plots the average proportion of optimal choices for participants best fit by each model. Participants who were best fit by the Equal Updating Softmax model (M=.78, N=15) made significantly more optimal choices than participants best fit by the Softmax model (M=.70, N=15), $F(1,28)=4.24$, $p<.05$, $\eta^2=.13$, making, on average, about 8% more optimal choices.

We expected that participants who were fit significantly better by the Equal Updating Softmax model would have significantly higher recency parameter values for the unchosen option (β) than participants fit best by the Softmax model. Figure 6b plots the average recency parameter values for the unchosen option for participants across all three conditions best fit by each model. As expected, participants who were best fit by the Equal Updating Softmax model (M=.83), had significantly higher average recency parameter values for the unchosen option than participants best fit by the Softmax model (M=.192), $F(1,28)=51.32$, $p<.001$, $\eta^2=.65$. This suggests that participants who altered their EV of the unchosen option based on more recent information about the chosen options outperformed those who did not do so.

Discussion

We were unable to account for the observed differences in performance across conditions in Experiment 2 using the standard two-parameter version of the Softmax model. However, we were able to provide evidence that more flexible updating of the

EVs of both the chosen and unchosen options led to more optimal performance on the task by adding another recency parameter to update the EV of the unchosen option. Participants who performed better on this task were generally better fit by the Equal Updating Softmax model. Participants in the Multiplied condition earned significantly more points on the task than participants in the Control condition who earned significantly more points on the task than participants in the Shifted condition. Likewise, participants in the Multiplied condition were fit better by the Equal Updating Softmax model than participants in the Control condition, and participants in the Control condition were fit better by the same model than participants in the Shifted condition. Although these differences in average model fits were not significant, there is still reason to believe that the advantage in performance among the three conditions on the behavioral measures was due to a difference in the speed at which participants in these conditions adjusted the EVs of both options based on the most recent information.

General Discussion

Our goal in conducting these experiments was to compare hypotheses derived by the structure of two common choice rules that are used in numerous models across a wide variety of domains. The reasons for using either the softmax or matching rule to compute a probability for action selection are seldom given, and the differences are rarely discussed (for exceptions see: Daw and Doya, 2006; Corrado et al., 2005). The differences between the two rules might be minor, and perhaps inconsequential in many applications of models using either of these rules. However, the differential output of the choice rules when EVs of varying ratio or distances are inserted is of importance to modelers and empirically-minded scientists alike.

In the current paper we highlighted the differences among the models, and then tested them in two studies. For participants in our Shifted conditions the distance between the EVs (ranging between 81 and 90 points) is the same as the distance between the EVs in the Control conditions (between 1 and 10 points), but the ratio between the EVs is much smaller. Indeed, as the absolute value of the EVs increases while the distance between them stays the same, the ratio between them asymptotes at .50. The matching rule then suggests that differences in EV should be less discriminable as the absolute value of the EVs increases. The results were consistent with this prediction. This pattern of data also fits with the negatively accelerated value function in Prospect Theory (Kahneman and Tversky, 1979). However, this pattern of data was inconsistent with the Softmax model which predicts that performance is driven by the distance between the expected value of the outcomes, rather than the ratio.

We found that participants receiving multiplied rewards between 10 and 100 points performed at least as well as those receiving rewards between 1 and 10 points, and sometimes significantly better. Their performance was also significantly better than participants in the Shifted condition. This result suggests that the inferior performance for participants in the Shifted condition was not simply due to the larger reward values that they had to process, because participants in the Multiplied condition had comparably large rewards. This result also supports the hypothesis extended from the softmax rule that as the distance between the EVs increases participants should exploit the option with the highest EV.

Our results suggest that there may be psychophysical differences in how humans process expected reward values. Rewards such as those given to participants in the

Shifted condition may be difficult to process for reasons other than the decrease in the ratio between the EVs. Another possible reason is simply that these rewards are more awkward to process than the rewards ranging from 1-10 points, or 10-100 points in increments of 10. One could argue that people have more experience processing the rewards given to participants in the Multiplied and Control conditions than those given to participants in the Shifted condition.

Individual Differences in Reward Processing

It is also important to point out that the failure of participants in the Shifted conditions to adequately exploit the decks with the highest EV was not uniform for all participants in those conditions. Indeed, there were some participants who seemed to have no trouble with this task, performing about as well as participants in the Control and Multiplied conditions, while others clearly had problems. There may be important individual difference variables that correlate with the ability to process values similar to those given to participants in our Shifted conditions, and performance on such a task may be predictive of superior or inferior performance on other tasks.

Several researchers have suggested that traits that characterize extraversion arise from differences in the sensitivity of the brain's reward system (Gray, 1970; Cohen, Young, Baek, Kessler, & Ranganath, 2005; Depue & Collins, 1999; Diener, Oishi, & Lucas, 2003) In a recent study, Cohen and colleagues (2005) observed that participants who scored higher on an extraversion personality questionnaire exhibited a greater reward response in the Orbitofrontal cortex (OFC) and nucleus accumbens which were shown to be reward sensitive areas. However, they failed to find any behavioral correlates with level of extraversion.[3] Future work should be done to identify the extent to which extraversion predicts reward-sensitivity in choice tasks.

There also may be individual differences found for participants who are best fit by the Equal Updating Softmax model that we developed to analyze Experiment 2. An implication of this model is that both the chosen and unchosen options are updated after receiving each reward, and that the reward earned for choosing one option influences the EV of the other option. This idea may have implications extending to domains such as business and marketing where a poor experience with one product or service leads to an improved view of others. Individual differences or factors that influence the degree to which one changes their views of unchosen options, products, or services may be of interest to companies looking to provide an alternative choice for the consumer.

The Neuroscience of Reward Processing

These results also relate to several avenues of research on the neurobiology of reward processing. Several studies have implicated the OFC (Schultz, 2000; Wallis and Miller, 2003; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001), the mesial prefrontal cortex (Knutson, Fong, Bennet, Adams, & Hommer, 2003) and the nucleus accumbens (Knutson, Adams, Fong, & Hommer, 2001) in the processing of reward. However, much of this work has been conducted with animals, and the type of symbolic reward processing (i.e. processing numbers for an ultimate goal of earning money) used in our task may be different than the reward processing occurring in other primates.

[3] In Experiment 2 participants' level of Extraversion was measured before the Experiment using Eysenck's Personality Questionnaire (Eysenck & Eysenck, 1975). Although, there were no significant correlations between level of Extraversion and the Total Adjusted Points earned on the task, participants in the Shifted condition did have a higher correlation coefficient (.21) between these measures than participants in the Control (0) and Multiplied (.03) conditions.

Imaging experiments may help us better understand the brain regions involved in reward processing.

Conclusion

Modeling choice and action selection is quite a complex task. Although simple models like the Softmax and Matching models can often predict a good deal of behavior, they are nevertheless too simple to account for all the complexities that people exhibit when trying to decide which option to pick to maximize reward. Here we have shown how two popular action selection rules make contrasting theoretical predictions about choice behavior, and that the predictions of each rule diverge from people's performance under some conditions. The important result here is that the reward values used in choice experiments must be chosen carefully. Both humans and learning models process rewards of varying magnitudes quite differently, and these differences should not be simply ignored.

References

Busemeyer, J. R. & Stout, J. C. (2002). A Contribution of Cognitive Decision Models to Clinical Assessment: Decomposing Performance on the Bechara Gambling Task. *Psychological Assessment*, 14, 253-262.

Cohen, M.C., & Massaro, D.W. (1992). On the Similarity of Categorization Models. In F.G. Ashby (Ed.), Multidimensional Models of Perception and Cognition. (pp. 395-448). Erlbaum: Hillsdale, NJ.

Cohen, M.X., Young, J., Baek, J.M., Kessler, C., Ranganath, C. (2005). Individual differences in extraversion and introversion and dopamine genetics predict neural reward responses. *Cognitive Brain Research,* 25, 851-861.

Corrado, G.S., Sugrue, L.P., Seung, S.H., & Newsome W.T. (2005). Linear-nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior*, 84, 581-617.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical Substrates for exploratory decisions in humans. *Nature,* 441 (15), 876-879.

Depue, R.A., Collins, P.F. (1999). Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences,* 22, 491-517.

Diener, E., Oishi, S., & Lucas, R.E. (2003). Personality, culture, and subjective well-being: emotional and cognitive evaluations of life. *Annual Review of Psychology,* 54, 403-425.

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88, 848-881.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (Adults).* London: Hodder & Stoughton.

Gray, J.A. (1970). The psychophysiological basis of introversion—extroversion. *Behavioral Research and Therapy,* 8, 249-266.

Herrnstein, R.J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4, 267-272.

Knutson, B., Adams, C.M., Fong, G.W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of Neuroscience,* 21, RC159.

Knutson, B., Fong, G.W., Bennet, S.M., Adams, C.M., & Hommer, D. (2003). A region of the mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *NeuroImage,* 18, 263-272.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Lau, B., Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84, 555-579.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309-332.

Luce, R.D. (1959). *Individual Choice Behavior*. Wiley, NY.

Luce, R.D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189), New York: Wiley.

Maddox, W. T., & Ashby, F.G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics,* 53, 49-70.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, 85, 207-238.

Minsky, M., & Papert, S. (1968, 1988). *Perceptrons.* Cambridge, MA: MIT Press.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General,* 115, 39-57.

Nosofsky, R.M., & Palmeri, T.J. (1998). A rule-plus-exceptions model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin and Review, 5,* 345-369.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience,* 4, 95-102.

Reed, S.K. (1972). Pattern Recognition and Categorization. *Cognitive Psychology,* 3, 482-487.

Roberts, M. E., & Goldstone, R. L. (2006). EPICURE: Spatial and Knowledge Limitations in Group Foraging. *Adaptive Behavior*, *14*, 291-313.

Rumelhart, D. E., & McClelland, J. L. (Eds.) (1986). *Parallel distributed processing, Vol. 1: Foundations.* Cambridge: MIT press.

Schultz, W. (2000). Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology*, 14, 139–147.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika,* 22, 325-345.

Sugrue, L.P., Corrado, G.S., Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, 304:1782-1787.

Sutton, R.S., & Barto, A.G. *Reinforcement Learning: An Introduction* MIT Press, Cambridge, Massachusetts, 1998.

Wallis, J.D., & Miller, E.K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience,* 18, 2069-2081.

Wickens, T.D. (1982). *Models for behavior: Stochastic processes in psychology.* San Francisco: W. H. Freeman.

Williams, R. J. (1988). On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, 263-270, San Diego, CA.

Worthy, D.A, Maddox, W.T., & Markman, A.B. (2007). Regulatory Fit Effects in a Choice Task. *Psychonomic Bulletin and Review,* 14, 1125-1132.

Yechiam, E., Busemeyer, J.R, Stout, J.C, and Bechara, A. (2005) Using cognitive models to map relations between neuropsychological disorders and human decision making deficits. *Psychological Science*, 16, 973-978.

Acknowledgements

This research was supported by AFOSR grant FA9550-06-1-0204, NIMH grant MH077708 to WTM and ABM, and a supplement to NIMH grant MH077708 to DAW. We thank Scott Lauritzen, Bo Zhu, Samaneh Pourali, Samir Zughni, Alex Hunt, Sahna Reddy, Sandeep Pinnapureddy, and Krystal Lopez for help in collecting the data.

Correspondence concerning this article should be addressed to W. Todd Maddox or Arthur Markman both at: University of Texas, 1 University Station A8000, Department of Psychology, Austin, Texas, 78712 (e-mail: maddox@psy.utexas.edu or markman@psy.utexas.edu.)

Appendix 1

Matching rule.

It can be shown that for two choice alternatives the matching rule reduces to the ratio between the two alternatives.

Let *x* represent the EV of choice a, *y* represent the EV of choice b, and r represent the ratio between *x* and *y*.

Let

$$r = \frac{x}{y}$$

$$y = rx$$

The matching rule can then be stated as

$$P_{a,t} = \frac{x}{x + rx}$$

$$= \frac{x}{(1+r)x}$$

$$= \frac{1}{(1+r)}$$

So that x cancels out and only the ratio of x and y remains.

Softmax rule

Similarly, it can be shown that the softmax rule reduces to the distance between the two alternatives.

Let *x* represent the EV of choice a, *y* represent the EV of choice b, and *d* represent the distance between *x* and *y*.

Let

$$d = y - x$$

$$y = d + x$$

The softmax rule can then be stated as:

$$P_{a,t} = \frac{e^x}{e^x + e^{(d+x)}}$$

$$= \frac{e^x}{e^x + (e^x * e^d)}$$

$$= \left(\frac{e^x}{e^x + e^x}\right)\left(\frac{1}{1 + e^d}\right)$$

$$= \left(\frac{1}{2}\right)\left(\frac{1}{1 + e^d}\right)$$

So that *x* cancels out and only the distance between *x* and *y* remains.

Table 1

Adjusted Average Exploitation Parameter Estimates for Participants in Experiment 1

Model	Matching	Softmax
Control	4.13 (.53)	0.84 (.09)
Shifted	3.38 (.63)	0.49 (.09)
Multiplied	4.42 (.67)	0.90 (.13)

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

Table 2

Adjusted Average Exploitation Parameter Estimates for Participants in Experiment 2

Model	Matching	Softmax
Control	2.56 (.30)	0.61 (.07)
Shifted	2.65 (.75)	0.61 (.17)
Multiplied	3.22 (.46)	0.72 (.09)

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

Figure Captions

Figure 1. (a) Average total adjusted points earned by participants in each condition in Experiment 1. (b) Average proportion of optimal choices made by participants in each condition in Experiment 1.

Figure 2. (a) Average total adjusted points earned by participants in each condition in Experiment 2. (b) Average proportion of optimal choices made by participants in each condition in Experiment 2.

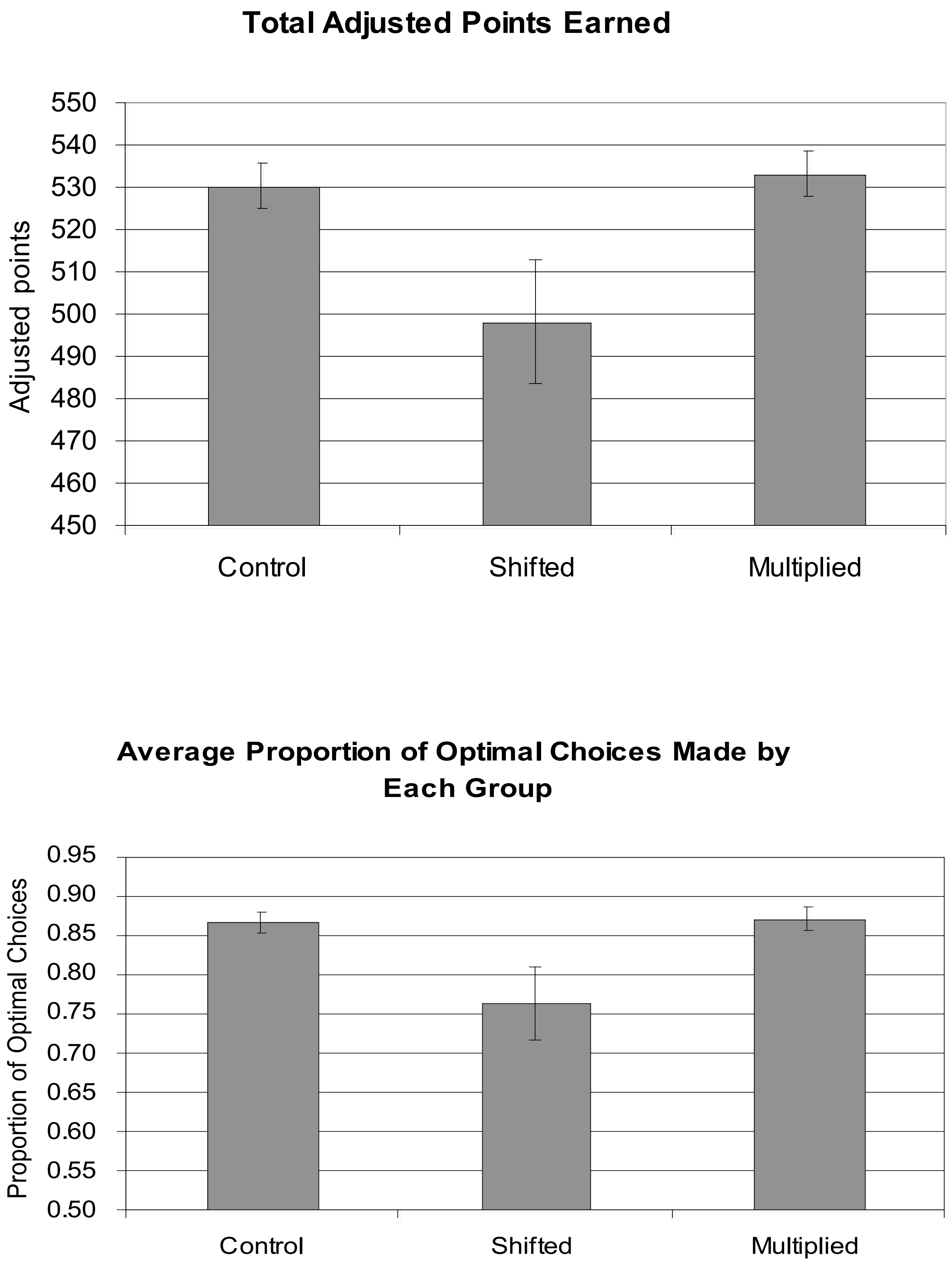
Figure 3. Proportion of participants best fit by the Softmax model in each condition of Experiment 1.

Figure 4. Average recency parameter values for the unchosen option estimated by the Equal Updating Softmax model for participants in Experiment 2.

Figure 5. Average G^2 fit values estimated by the Equal Updating Softmax model for participants in Experiment 2.

Figure 6. (a) Average proportion of optimal choices made by participants in Experiment 2 who were fit best by either the Softmax or the Equal Updating Softmax model. (b) Average estimated recency parameter values for the unchosen option estimated by the Equal Updating Softmax model for participants in Experiment 2 who were best fit by either the Softmax or the Equal Updating Softmax model.

Figure 1



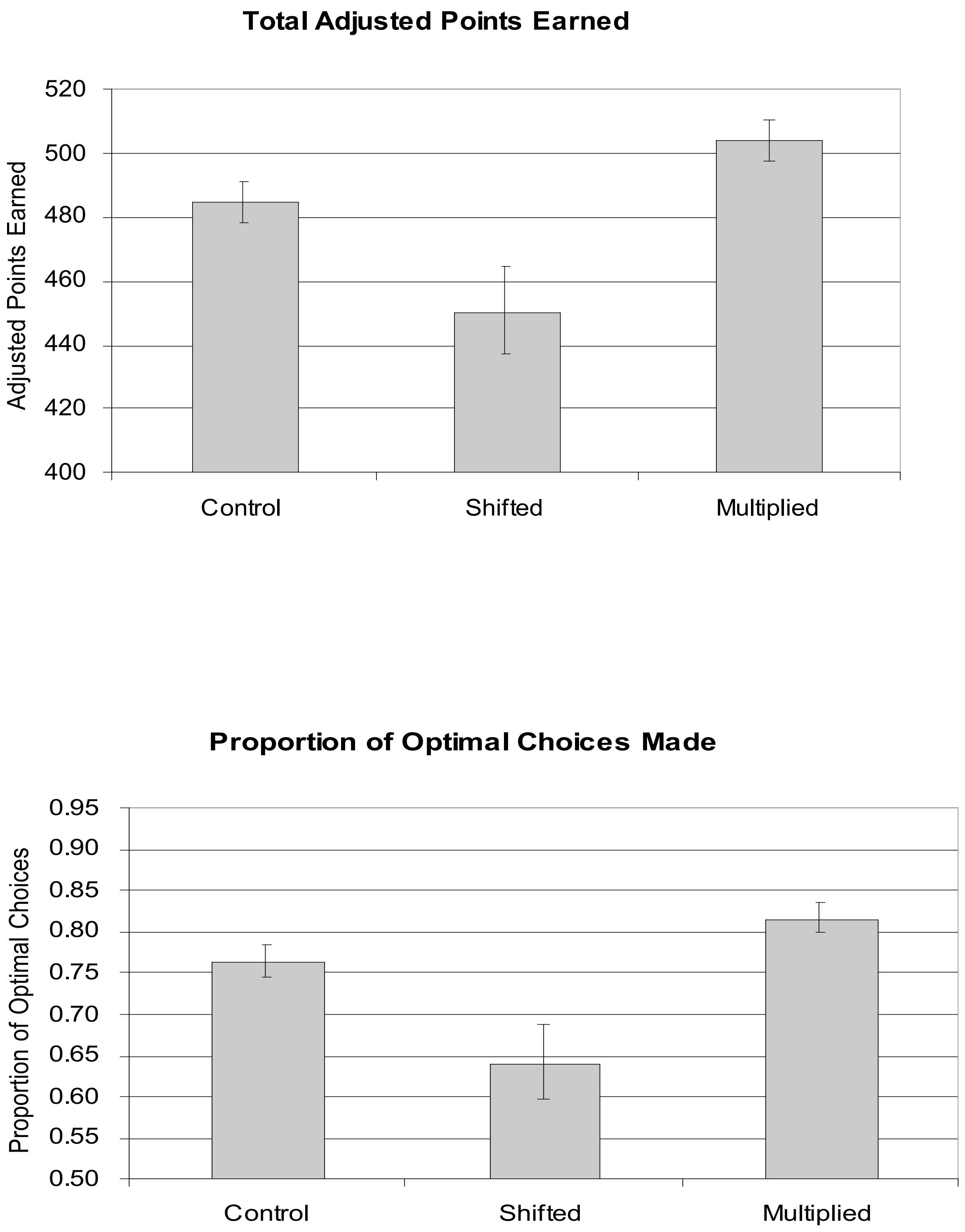
a.

Total Adjusted Points Earned

b.

Average Proportion of Optimal Choices Made by Each Group

Figure 2



a.

Total Adjusted Points Earned

b.

Proportion of Optimal Choices Made

Figure 3

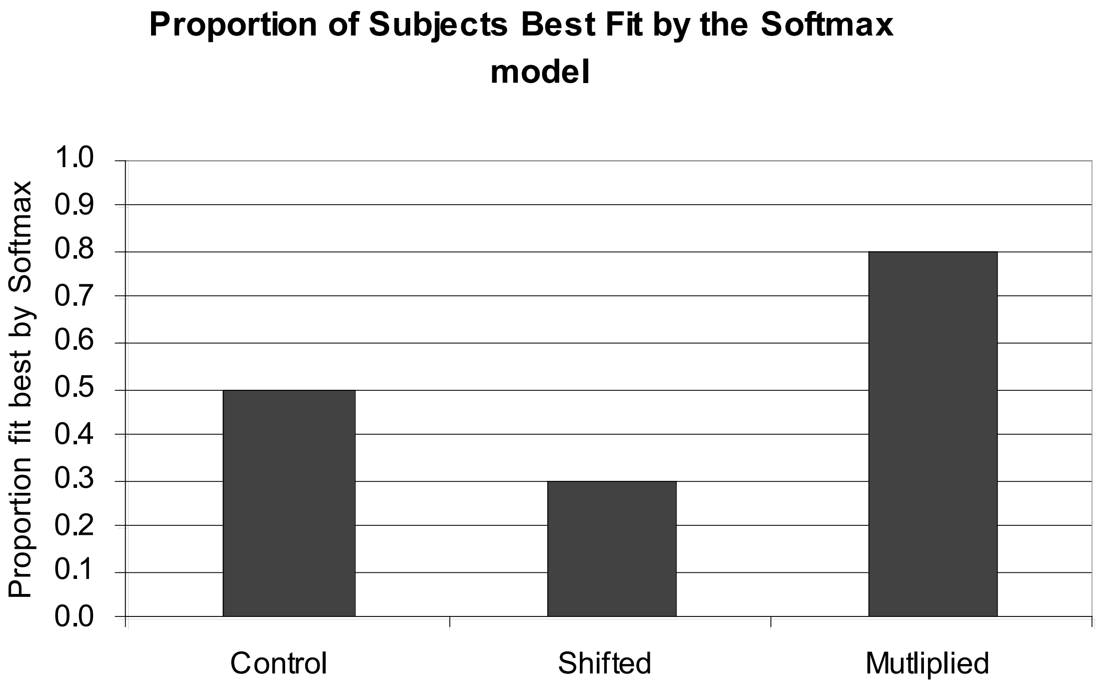

Proportion of Subjects Best Fit by the Softmax model
Proportion fit best by Softmax
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Control
Shifted
Mutliplied

Figure 4

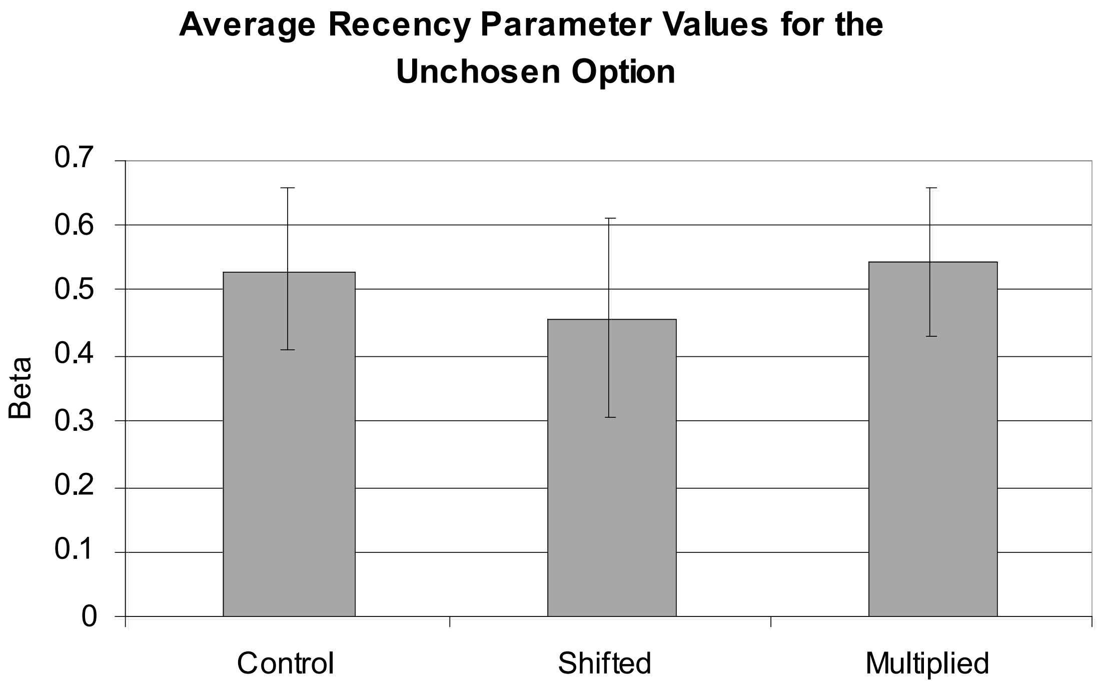
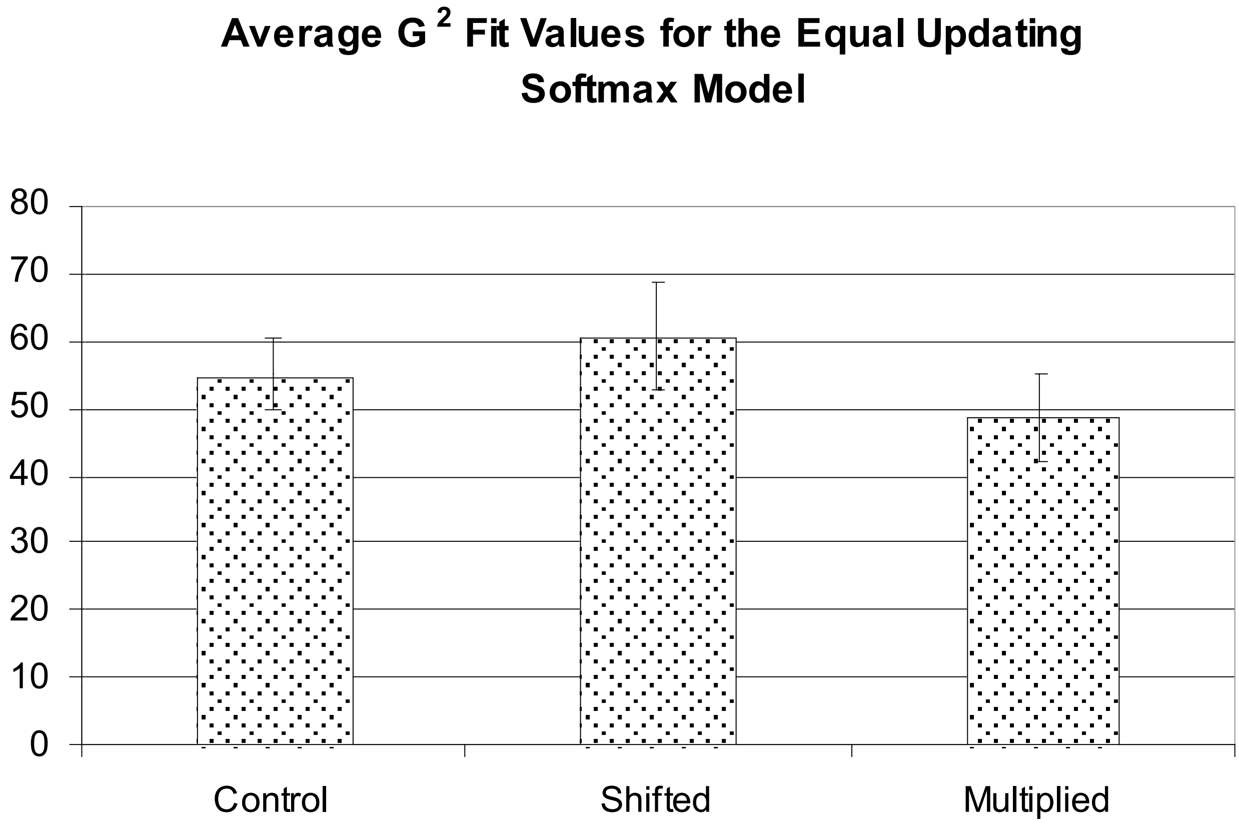
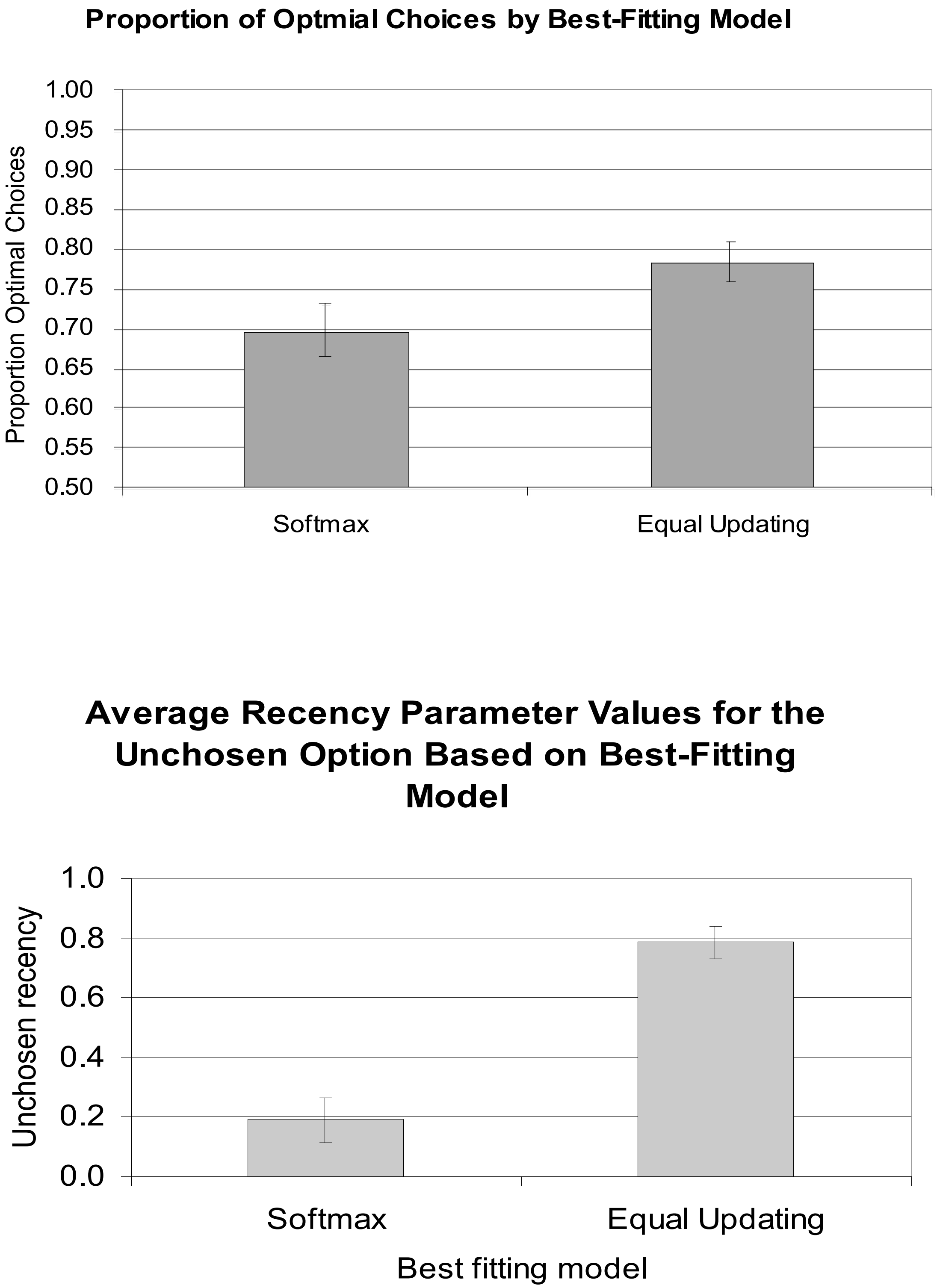

Average Recency Parameter Values for the Unchosen Option
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Beta
Control
Shifted
Multiplied

Figure 5



Average G^2 Fit Values for the Equal Updating Softmax Model

Figure 6



a.

Proportion of Optmial Choices by Best-Fitting Model

b.

Average Recency Parameter Values for the Unchosen Option Based on Best-Fitting Model