

Processing Expected Reward in Decision-Making Tasks

Darrell Worthy

Department of Psychology

University of Texas at Austin

Collaborators: Todd Maddox and Art Markman

Outline

- The n -armed bandit problem
 - exploration vs. exploitation
 - modeling choice behavior
- Comparison of Expected Values (EVs)
 - ratios vs. absolute differences
- Theoretical vs. Descriptive Uses of Models
- Experimental tests (from model predictions)
- Discussion
- Implications

Introduction

- Decision making paradigm
- n -armed bandit task
 - Series of choices from n options
 - Gain points on each trial
 - Goal is to maximize gains for monetary reward

e.g. Bechara et al., 1994; Estes, 1950; Bush and Mosteller, 1955; Yechiam et al., 2005; Worthy et al., 2007; see also Sutton and Barto, 1998

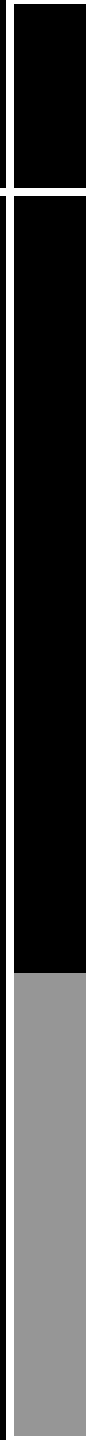
PICK A CARD!

Yes
Bonus
No

450

174

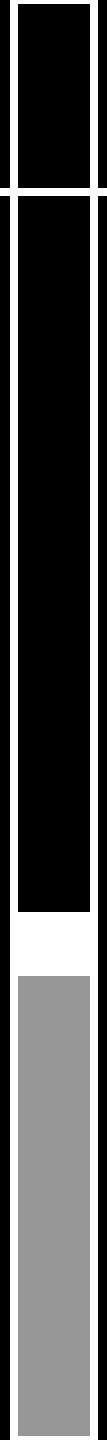
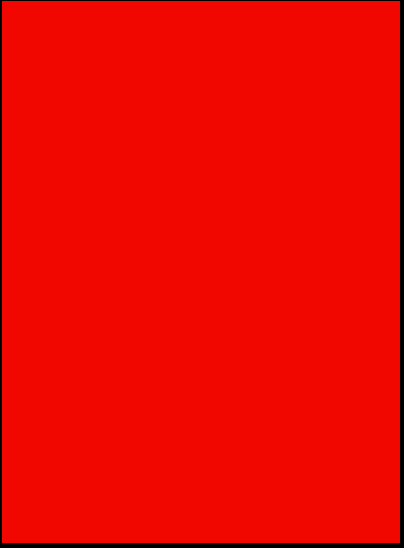
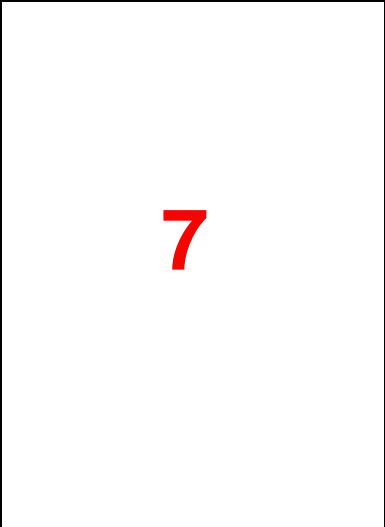
0



Yes
Bonus
No

450

181
174



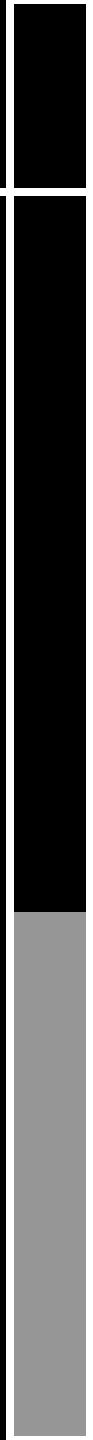
PICK A CARD!

Yes
Bonus
No

450

181

0



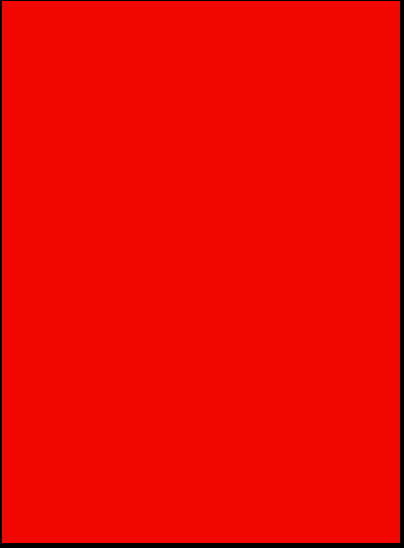
Yes
Bonus
No

3

450

184

181



PICK A CARD!

Yes
Bonus
No

450

184

0



Expected Value (EV)

- EV – How many points you expect to earn from selecting a given deck
- Used to determine which option to choose
- Example
 - $EV_{\text{red deck}} = 7$ points
 - $EV_{\text{blue deck}} = 3$ points

Exploitation/Exploration Dilemma

- *Exploit* the option with the highest EV
 - *or*
 - *Explore* other options with lower EVs
-
- Must balance the need to exploit with the need for new information

Modeling Choice Behavior

- EVs of each option are updated via an exponential recency-weighted algorithm

$$\underset{\substack{\uparrow \\ \text{New EV}}}{EV}_{k+1} = \underset{\substack{\uparrow \\ \text{Current EV}}}{EV}_k + \underset{\substack{\uparrow \\ \text{Recency} \\ \text{Parameter}}}{\alpha} \left[\underset{\substack{\uparrow \\ \text{Reward}}}{r}_{k+1} - \underset{\substack{\uparrow \\ \text{Current EV}}}{EV}_k \right]$$

- If reward is greater than the current EV the EV increases
- If reward is less than the current EV the EV decreases

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

- α is a free parameter constrained to be between 0 and 1
- Higher α values give greater weight to recent rewards
- When $\alpha = 1$, Updating Equation reduces to:

$$EV_{k+1} = r_{k+1}$$

- Alternatively, when $\alpha = 0$, Updating Equation reduces to:

$$EV_{k+1} = EV_k$$

Action Selection

- Action selection is probabilistically determined via choice rules (e.g. Luce, 1959)

Softmax Rule

Probability of choosing option "A" $\rightarrow P_{a,t} = \frac{e^{(\gamma EV_t(a))}}{\sum_{b=1}^n e^{(\gamma EV_t(b))}}$

Exploitation parameter $\leftarrow \gamma$

EV for option "A" $\leftarrow EV_t(a)$

Sum of EVs for all options $\leftarrow \sum_{b=1}^n e^{(\gamma EV_t(b))}$

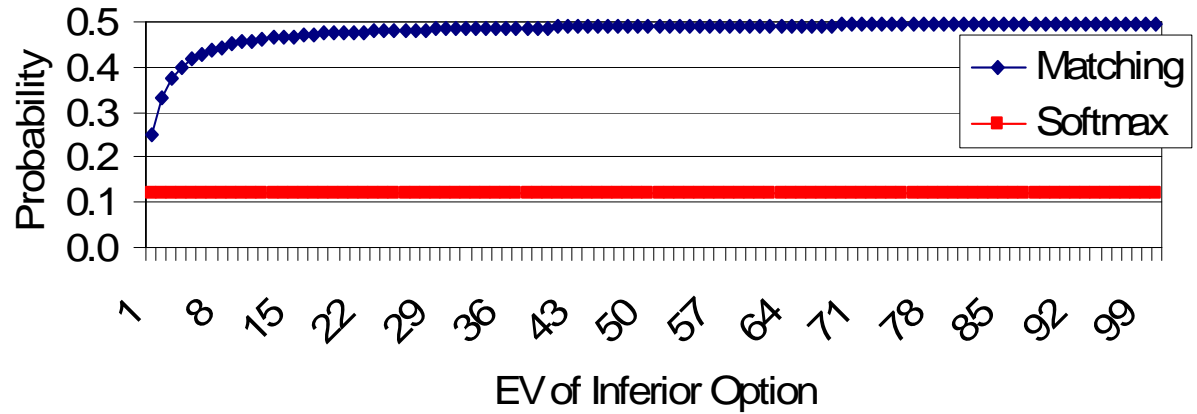
- Higher γ values indicate greater exploitation
- Lower γ values indicate greater exploration

Matching Rule

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

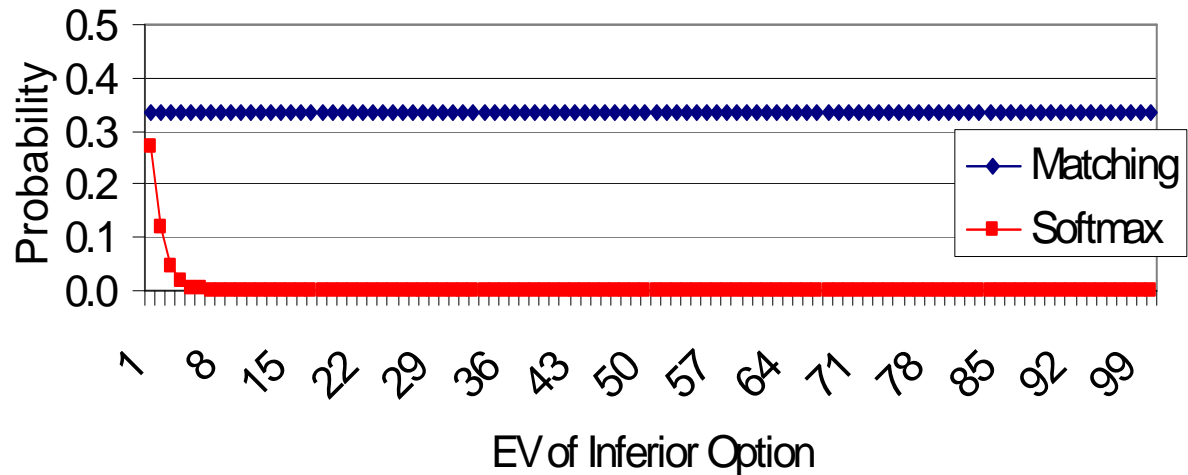
- Non-exponentiated form of the Softmax Rule
- Softmax Rule compares the *difference* between EVs
- Matching Rule compares the *ratio* between EVs

CONSTANT DIFFERENCE: Action Probabilities for an Option with an EV 2 Units Less than the Alternative



Softmax Rule gives same probability when *distance* is preserved

CONSTANT RATIO: Action Probabilities for an Option with an EV that is Half the Value of the Alternative EV



Matching Rule gives same probability when *ratio* is preserved

- Various models use either distance dependent (Softmax) or ratio dependent (Matching) rules
- **Matching (Ratio-Dependent):**
 - Animal behavior (e.g. Herrnstein, 1961; Sugrue et al., 2004)
 - Category learning (e.g. Nosofsky, 1984; Nosofsky and Palmeri, 1998; Maddox and Ashby, 1993)
- **Softmax (Distance preserving):**
 - Choice behavior (e.g. Sutton and Barto, 1998; Daw et al., 2006)
 - Foraging behavior (e.g. Roberts and Goldstone, 2005)
 - Category learning (e.g. Kruschke, 1992; Love et al., 2004)

Reasons for Using Either Rule

- Rarely discussed
- Rule used may be arbitrary
- Frequently used rules make psychological predictions that could constrain future decision rules used.
- Tests of theoretical predictions of a model may be an alternative test of a model's validity to goodness-of-fit (e.g. Roberts and Pashler, 2000)
 - Current goal

Theoretical vs. Descriptive

- Models can be used to generate *theoretical* predictions about behavior, or to *describe* data (i.e. parameter estimates)
- Interested here in testing theoretical predictions generated from two common choice rules (Softmax and Matching).
- Test effects of preserving distance between EVs, varying ratio relative to control (Softmax Model); vice versa (Matching Model)
- Goal is not to say which is a “better” model but which model’s predictions are closer to behavior

Distance vs. Ratio

- Created three conditions to tests predictions: Control, Distance Preserving (alters ratio), Ratio Preserving (alters distance)

Control choice: $EV_t(a) = 7$; $EV_t(b) = 3$

Distance preserving: $EV_t(a) = 87$; $EV_t(b) = 83$

- The difference between the EVs is the same, ratio is different

Control Choice: $EV_t(a) = 7$; $EV_t(b) = 3$

Ratio Preserving: $EV_t(a) = 70$; $EV_t(b) = 30$

- The ratio between the EVs is the same, distance is different

- Softmax and Matching rules make different assumptions about decision making processes

Decision 1 (Control): $EV_t(a) = 7$; $EV_t(b) = 3$

Decision 2 (Distance): $EV_t(a) = 87$; $EV_t(b) = 83$

Decision 3 (Ratio): $EV_t(a) = 70$; $EV_t(b) = 30$

- **Softmax Rule Predicts:**

- Option *a* should be chosen *more* frequently in D3 than in D1 or D2 – greater *distance* = greater probability of exploitation
- There should be no difference between D1 and D2

- **Matching Rule Predicts:**

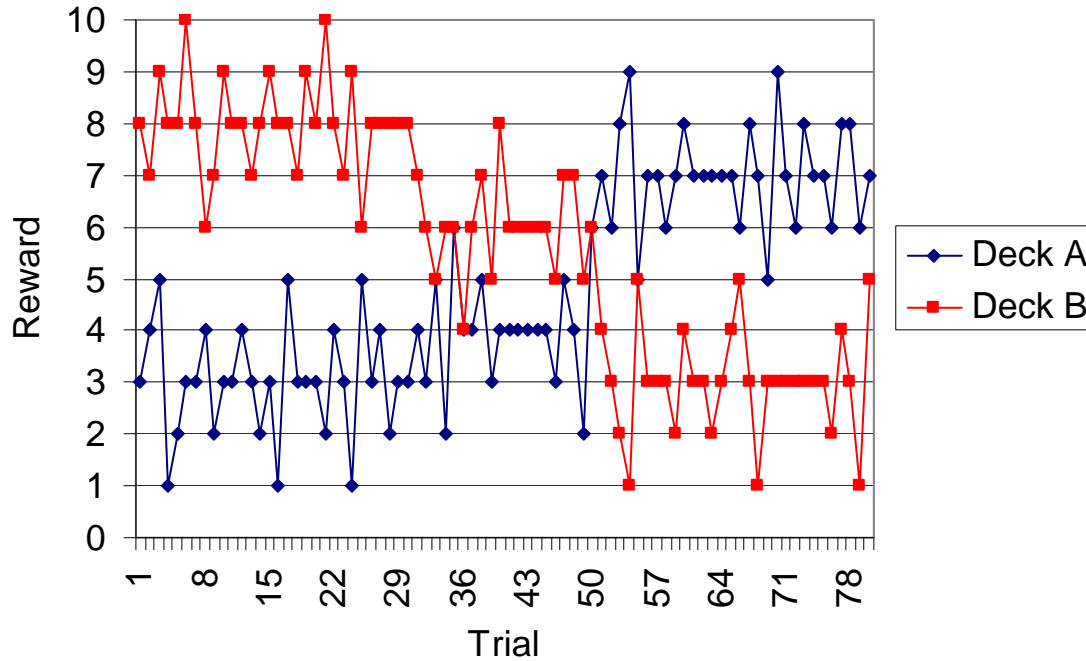
- Option *a* should be chosen *less* frequently in D2 than in D1 or D3 – smaller *ratio* = smaller probability of exploitation
- There should be no difference between D1 and D3

Experiment 1

- Tested the assumptions made by each rule with three conditions
- **Control**
 - All reward values were between 1 and 10 points
- **Distance Preserved**
 - Each reward from the Control condition shifted by 80
 - All reward values were between 81 and 90 points
- **Ratio Preserved**
 - Each reward from the Control condition multiplied by 10
 - All reward values were between 10 and 100 points
- Ten participants in each condition

Control Condition Design

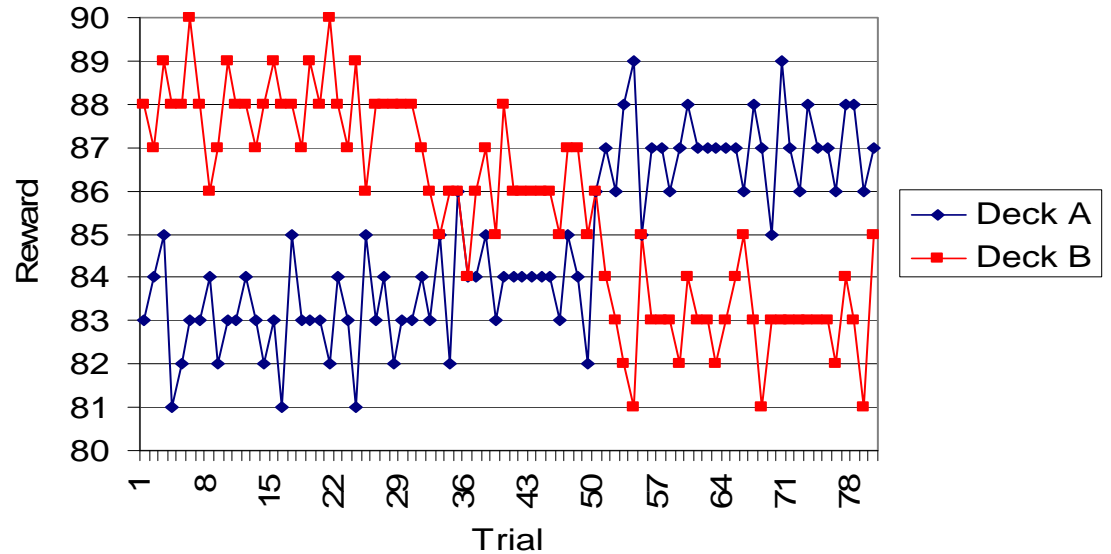
Reward Values for Control Condition



Deck A	Deck B
3 points over the first 30 trials	8 points over the first 30 trials
4 points over the next 20 trials	6 points over the next 20 trials
7 points over the last 30 trials	3 points over the last 30 trials

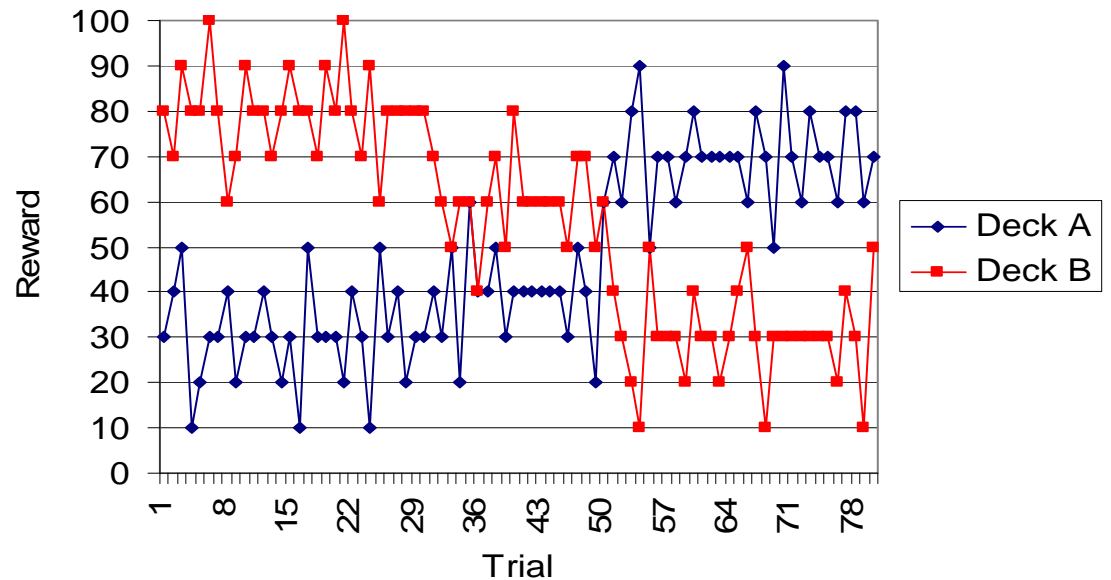
- 80 points added to each Control Value
- Preserves the Distance, Alters the Ratio

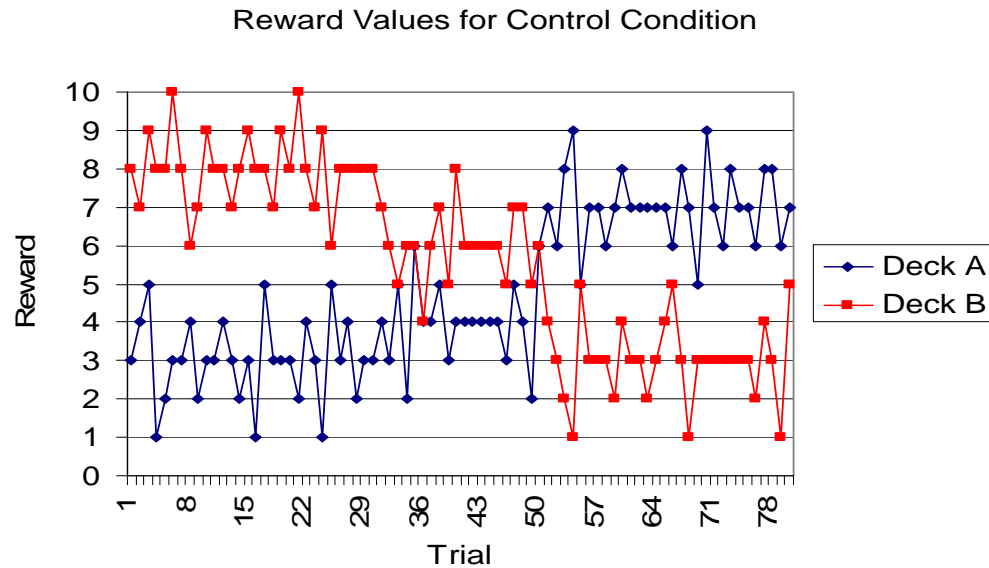
Reward Values for Distance Preserving Condition



- Each Control Value Multiplied by 10
- Preserves the Ratio, Alters the Distance

Reward Values for Ratio Preserving Condition

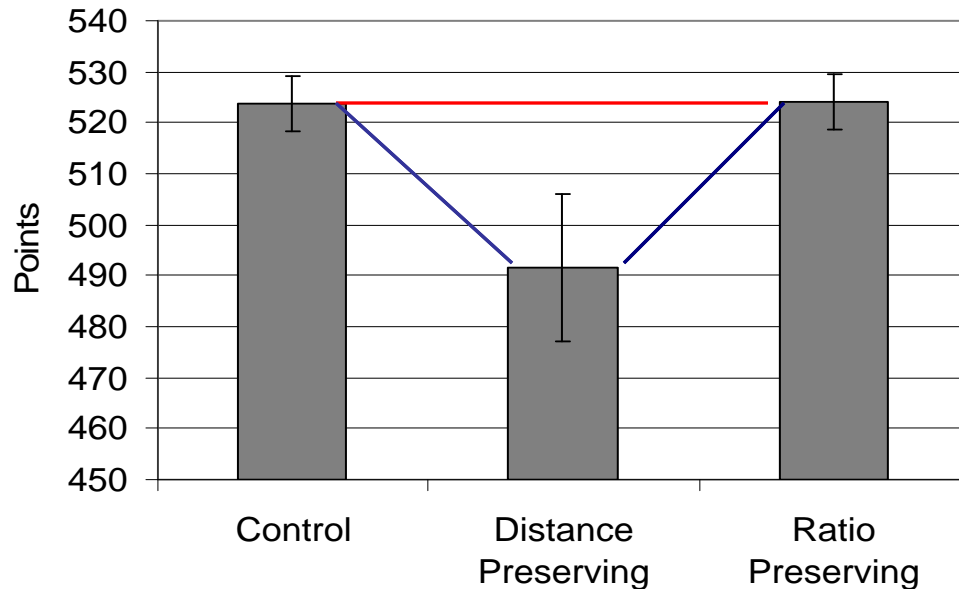




- Goal was to earn enough points to receive monetary bonus
- Control condition had to earn 550 points over 80 trials
 Shifted condition: $6400 + 550 = 6950$ points
 Multiplied condition: $550 * 10 = 5500$ points
- Optimal strategy was to exploit Deck B for the first 50 trials;
 Deck A for the final 30 trials

Results

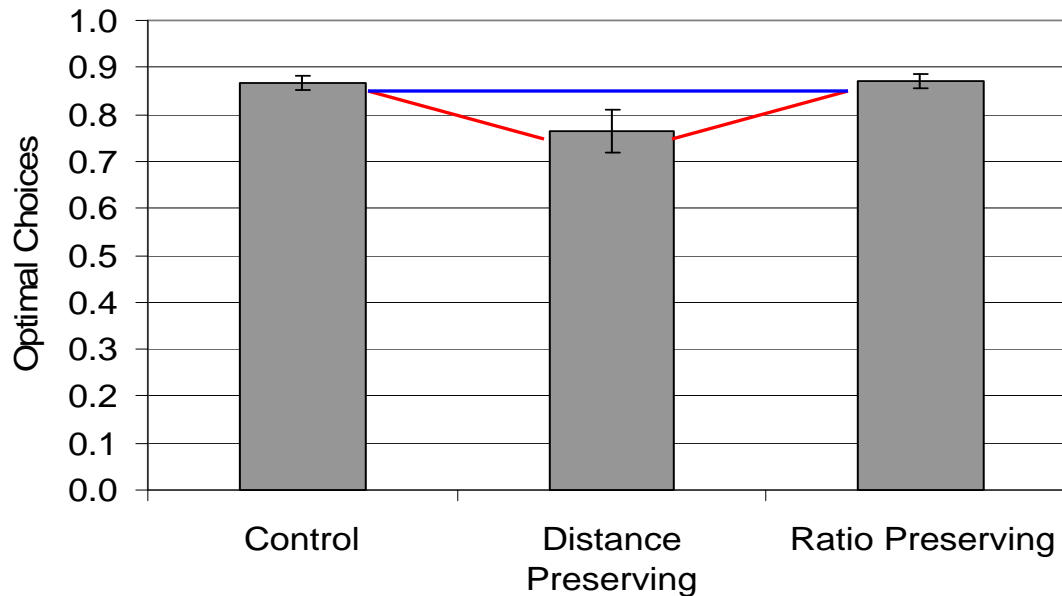
Total Adjusted Points Earned



- Participants in the Distance Preserving condition earned significantly fewer points than those in Control and Ratio Preserving conditions.
- No difference between Control and Ratio Preserving conditions
- Supports predictions from Matching Rule

Results

Proportion of Optimal Choices



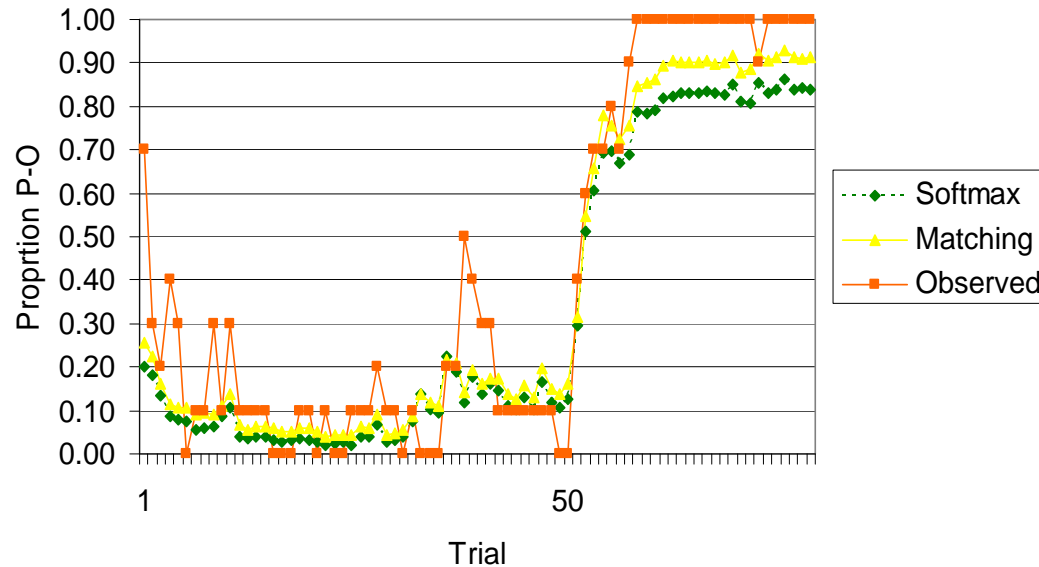
- Optimal choice = Deck B for trials 1-50; Deck A trials 51-80
- Participants in the Distance Preserving condition made significantly fewer points than those in Control and Ratio Preserving conditions
- No difference between Control and Multiplied conditions

Model-Based Analyses

- Fit Models with Softmax (Distance Dependent) and Matching (Ratio Dependent) rules
- Used same exponential recency-weighted algorithm to compute EV
- Fit each subject's data individually based on model's ability to predict the next choice
- Best fitting parameters using maximum log-likelihood.

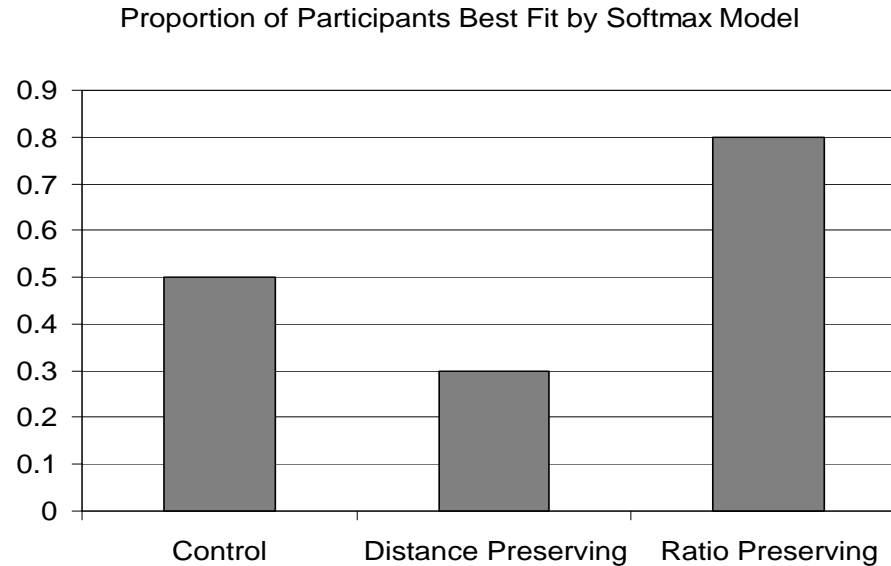
Model fits

Predicted and Observed Selections of Deck A for Control Group



- Fit each participant's data on a trial-by-trial basis
- Models generated a prediction for selecting either deck on each trial (Deck A shown here).
- Log likelihood of selected deck's selection probability summed over all trials

Model fits



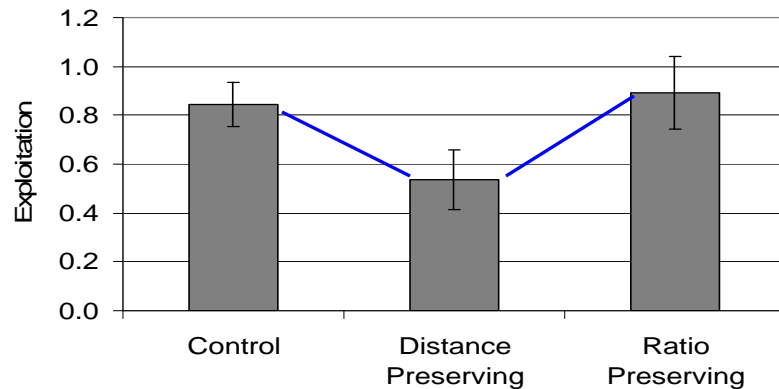
- More participants in the Distance Preserving condition best fit by the Matching model which depends on the ratio between EVs
- More participants in the Ratio Preserving condition best fit by the Softmax model which depends in the distance between EVs

Parameter Estimates

- Examined differences in Exploitation Parameter
- No differences in α -Recency parameter

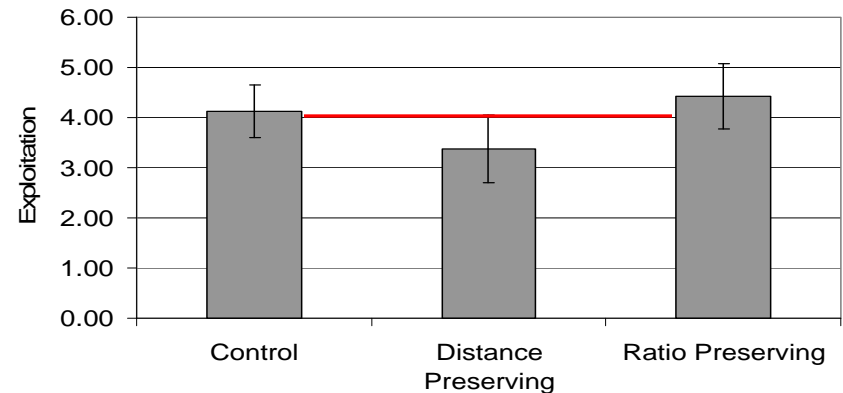
Softmax

Exploitation parameter values estimated by the Softmax model



Matching

Exploitation Parameter Values Estimated by the Matching Model



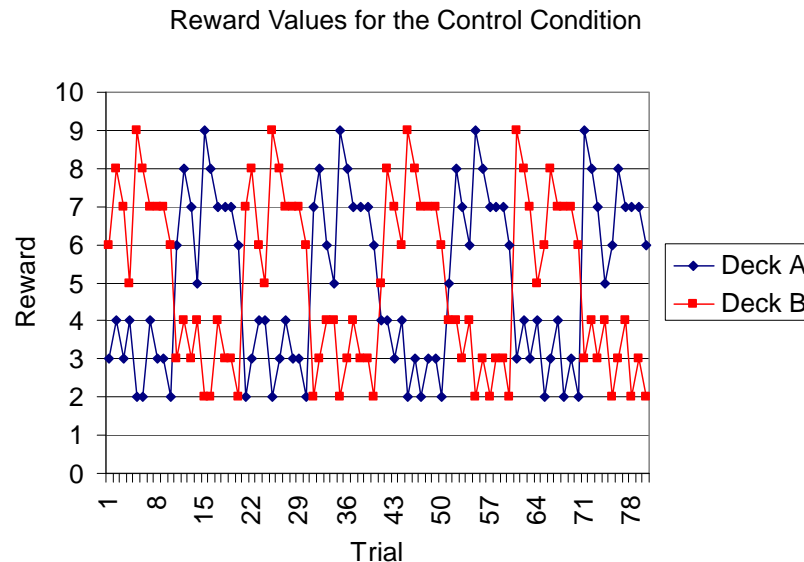
- Fits of Softmax model indicated greater exploitation for Control and Ratio Preserving conditions
- Interestingly, no significant differences for Matching model

Discussion

- Participants in the Distance Preserving condition had difficulty exploiting the option with the highest EV
- As Matching Rule predicts, as the ratio between EVs decreases so does the exploitation of best option
- Softmax Rule predicts no difference
- Ratio Preserving condition performed no better than Control
- Softmax model predicts advantage – greater distance
- Possible ceiling effect (optimal choices near 90 %)

Experiment 2

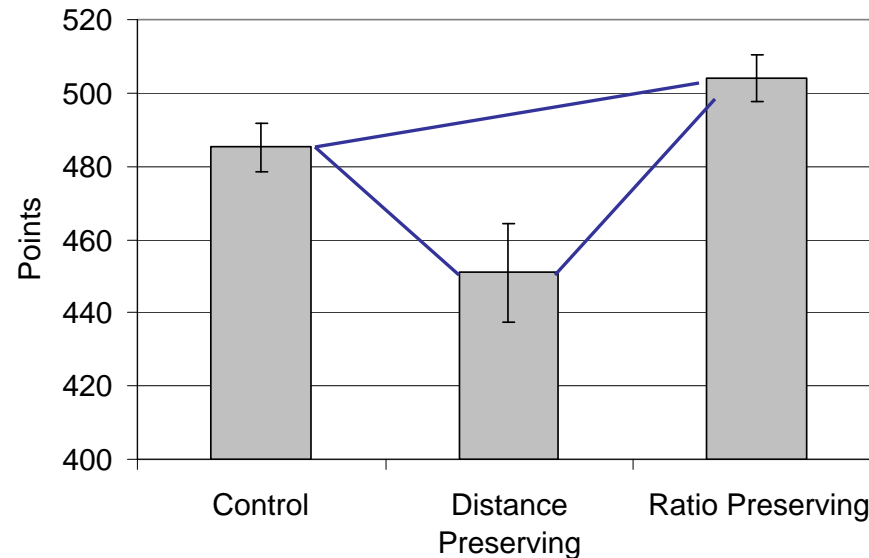
- Replicate deleterious effect of shifting rewards
- Increase difficulty by decreasing stationarity



- All other methods identical to Experiment 1

Results

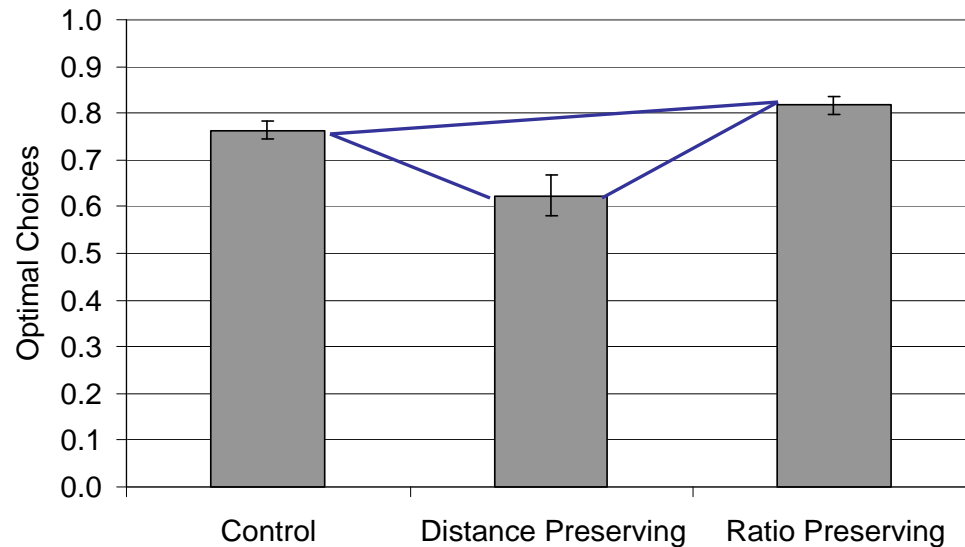
Total Adjusted Points Earned



- Significant differences between all three conditions with Ratio Preserving condition performing best.
- Supports Matching Rule's prediction of inferior Distance Preserving condition performance
- Supports Softmax Rule's prediction of superior Ratio Preserving condition performance

Results

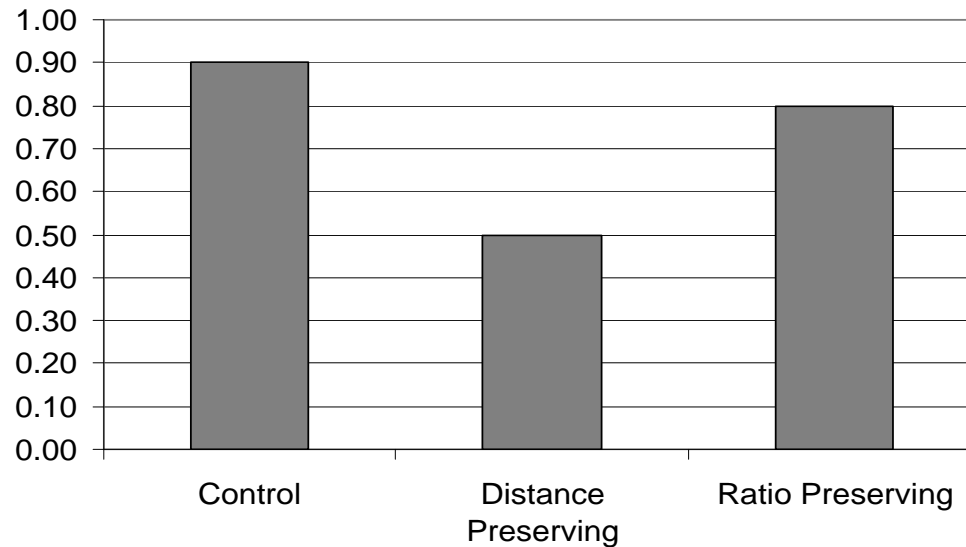
Proportion of Optimal choices



- Significant differences between all three conditions with Ratio Preserving condition performing best
- Corroborates Total Points data
- Proportion of optimal choices lower than in Experiment 1 (e.g Control condition 87 % vs. 76 % for Exps. 1 and 2)

Model fits

Proportion of Participants Best Fit by the Softmax Model



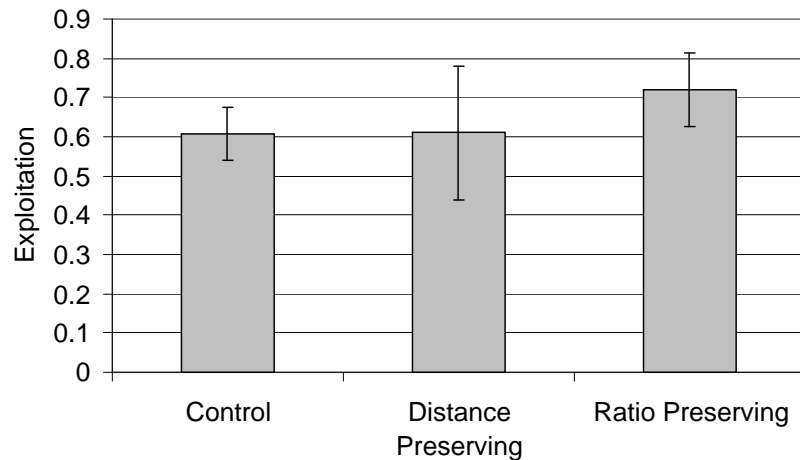
- More participants in Control and Ratio Preserving Conditions
- Equal number of participants fit by each model in Distance Preserving Condition

Model Based Analyses

- Again, no differences in α -Recency parameter – all were very high indicating greater weighting of recent rewards

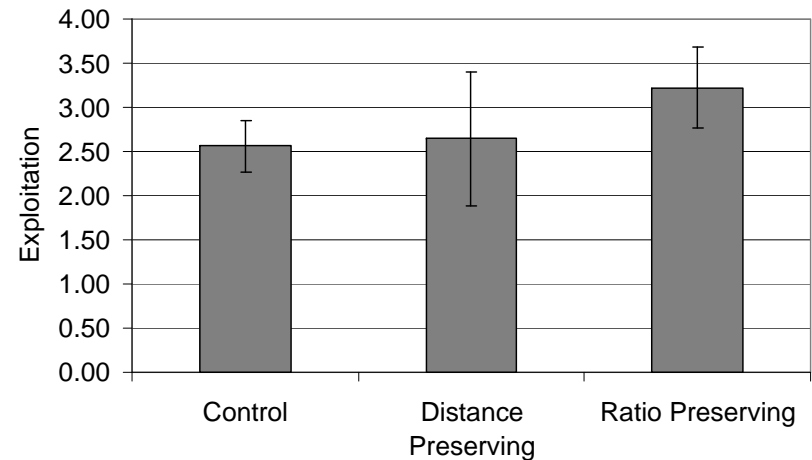
Softmax

Exploitation Parameter Values Estimated by the Softmax Model



Matching

Exploitation Parameters Estimated by the Matching Model



- Neither model can account for the observed behavioral differences
- May require a more complex model

Updating Unchosen Option

Chosen

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

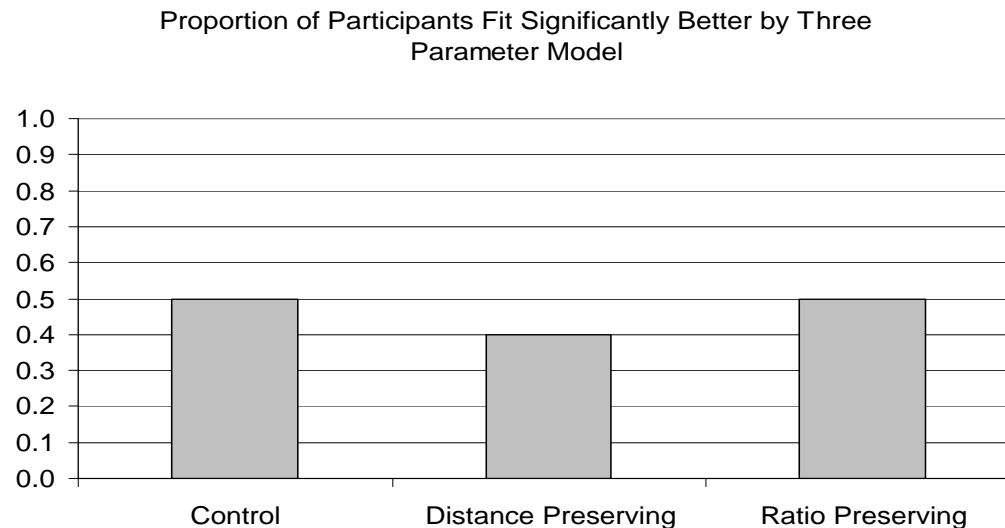
Unchosen

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

- Now unchosen option is updated based inversely on the reward for chosen option (e.g. if reward is less than expected other option's EV increases)
- Separate β Recency parameter determines influence of reward from chosen option
- Fit using Softmax Rule

Best Fitting Model

- Used chi-square tests for nested models to determine if adding β recency parameter significantly improved fit.

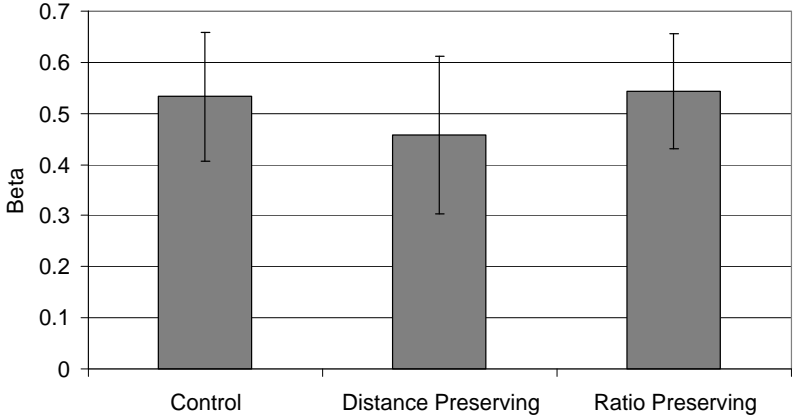


- About half of participants fit significantly better by adding β recency parameter.

β Recency

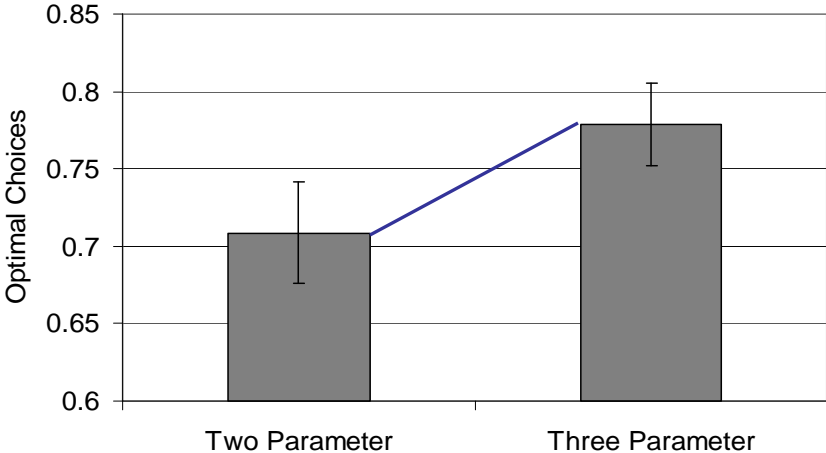
- No differences between groups

Estimated Recency Parameter Values for Unchosen Option



- However, participants across all groups who were fit better by adding β parameter made a higher proportion of optimal choices.

Proportion of Optimal Choices Based on Best-Fitting Model



Discussion

- As predicted by the Matching Rule, *decreased* ratios between EVs leads to *decreased* exploitation of the best option.
- Found support in Experiment 2 for the Softmax Rule's prediction that *increased* absolute distance between EVs leads to *increased* exploitation of the best option.

- Suggests psychophysical differences in how humans process expected reward value
- Inferior performance may not be due to ratio comparisons
- Shifted rewards may simply be more awkward
- Less experience with rewards from Shifted condition

Implications

- Modelers often ignore differences in choice rules
 - Whether model sums or averages evidence may alter choice probabilities (e.g. summed similarities vs. incremental update average)
- Theoretical vs. Descriptive Value of Models
 - Matching model's theoretical predictions were better supported, but parameter estimates from Softmax model accounted for behavioral differences better.
- May inform neural models of reward processing
 - Different systems for comparing ratios and differences
 - Predispositions for processing certain types of reward

Implications

- Ratio and difference comparisons in other domains
 - Comparisons for weight, pitch, loudness, darkness etc.
 - Fechner-Weber law
 - Stimulus intensity decreases as background noise increases
 - Too much background noise for Shifted condition
- Individual differences
 - Some subjects could exploit good decks in Shifted condition, others could not

Future Directions

- Develop or implement a more sophisticated learning model for non-stationary environments as in Experiment 2
- Explore individual differences in reward processing
- Examine neural correlates of reward
- Relate the current work to previous work in psychophysics on ratio and difference comparisons (e.g. Fechner-Stevens etc.)

Acknowledgements

Thanks to Todd Maddox and Art Markman; Scott Lauritzen and Maddox lab undergraduate research assistants for help with data collection; Markman lab and Motivation group for helpful comments. Supported by AFOSR FA9550-06-1-0204 to WTM and ABM.

References

- Bechara A., Damasio A.R., Damasio H., Anderson S.W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7–15.
- Bush, R. R. and Mosteller, F. (1955). *Stochastic Models for Learning*. Wiley, New York.
- 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.
- Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, 57:94-107.
- 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.
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.
- Maddox, W. T., & Ashby, F.G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53, 49-70.
- 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.
- Roberts, M. E., & Goldstone, R. L. (2006). EPICURE: Spatial and Knowledge Limitations in Group Foraging. *Adaptive Behavior*, 14, 291-313.
- Roberts S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358-367.
- 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.
- 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.