Influence of learning strategy on response time during complex value-based learning and choice

Abbreviated title: Construction of reward value and response time

Shiva Farashahi[1], Katherine Rowe[1], Zohra Aslami[1], M Ida Gobbini[1,2], Alireza Soltani[1*]

[1]*Department of Psychological and Brain Sciences, Dartmouth College, NH, United States of America*
[2] *Dipartimento di Medicina Specialistica, Diagnostica e Sperimentale (DIMES), Medical School, University of Bologna, Bologna, Italy*

Correspondence: A.S., Department of Psychological and Brain Sciences, HB 6207, Dartmouth College, Hanover, NH, 03755. Email: soltani@dartmouth.edu

Manuscript information: 23 pages; 6 figures; and 3 tables.
Abstract 210 words
Introduction 680 words
Discussion 915 words

Conflict of interest: The authors declare no competing interests.

Acknowledgments: This work was supported by NSF EPSCoR (RII Track-2 FEC) grant to A.S.

January 15, 2017

Abstract

Measurements of response time (RT) has long been used to infer neural processes underlying various cognitive functions such as working memory, attention, and decision making. However, it is currently unknown if RT is also informative about various stages of value-based choice, particularly how reward values are constructed. To investigate this, we analyzed the pattern of RT during a set of multi-dimensional learning and decision-making tasks that can prompt subjects to adopt different learning strategies. In our experiments, subjects could use reward feedback to directly learn reward values associated with alternative choice options (object-based learning). Alternatively, they could learn reward values of options' features (e.g. color, shape) and combine these values to estimate reward values for individual options (feature-based learning). We found that RT was slower when the difference between subjects' estimates of reward probabilities for the two alternative objects on a given trial was smaller, and when the preceding trial was not rewarded. More importantly, RT reflected the learning strategy adopted by the subject on a trial-by-trial basis, indicating an overall faster construction of reward value and/or value comparison during object-based learning. Altogether, these results demonstrate that the pattern of RT can be used to infer how reward values are learned and constructed during complex value-based learning and decision making.

Introduction

Investigations of the response time (RT) have long been the focus of human psychophysics studies in order to distinguish between alternative mechanisms underlying mental processes (Donders, 1868; Luce, 1986; Sternberg, 1969b). Similarly, psychologists and neuroscientists have used the pattern of RT in different experimental conditions to infer neural processes underlying various cognitive functions. These include but are not limited to: visual processes (Breitmeyer, 1975; Duncan & Humphreys, 1989; Tanaka & Shimojo, 1996; Treisman & Sato, 1990; Tynan & Sekuler, 1982), attention (Joseph, Chun, & Nakayama, 1997; Krummenacher, Grubert, & Müller, 2010; Krummenacher, Müller, & Heller, 2001; Treisman & Sato, 1990; Wolfe, Cave, & Franzel, 1989), memory (Sternberg, 1969a; Unsworth & Engle, 2005), and decision making (Logan, Cowan, & Davis, 1984; Pleskac & Busemeyer, 2010; Ratcliff, Van Zandt, & McKoon, 1999; Spiliopoulos & Ortmann, 2016). Here we asked whether RT could also be used to infer how reward values are learned and constructed in the brain.

In naturalistic settings, reward could be associated with different aspects or features of an action or of an object. In order to be adaptive, learning from reward feedback needs to be adjusted according to the structure of this association in each environment. This learning, however, can be very challenging since options have many features, each of which can take different values, resulting in a large number of options for which reward values have to be learned (curse of dimensionality). Currently, there is an extensive literature on how humans exploit various heuristics to tackle real-world decision-making and learning tasks (Fellows, 2006; Gigerenzer & Goldstein, 1996; Gilovich, Griffin, & Kahneman, 2002; Jocham et al., 2016; Niv et al., 2015). We have recently shown that humans use a heuristic strategy to tackle learning reward values in dynamic multi-dimensional environments (Farashahi, Rowe, Aslami, Lee, & Soltani, 2017). More specifically, rather than directly estimating the reward value of individual options based on reward feedback (object-based learning), subjects estimate the average reward values of features and then combine these values to estimate the reward values for individual options (feature-based learning). Importantly, these learning strategies not only require different representation of reward values and updates based on reward feedback but also involve distinct methods for construction or comparison of reward values to make a decision.

We hypothesized that the way reward values are constructed and compared could influence how quickly a choice between two alternative options can be made. To examine this influence, we considered three stages involved in value-based decision making. First, objects and/or their features need to be detected and identified in order to evoke/elicit the corresponding representation of reward values. Second, in the case of feature-based learning, the individual values of relevant features should be combined to construct the final value of an object. Third, the final values of two alternative objects should be compared to make a decision (Lee, Seo, & Jung, 2012; Soltani & Wang, 2006, 2008; Sugrue, Corrado, & Newsome, 2005).

Because the choice during a given trial depends primarily on the difference between the reward values of the two options on that trial (Barraclough, Conroy, & Lee, 2004; Corrado, Sugrue, Seung, & Newsome, 2005; Lee et al., 2012; Louie & Glimcher, 2010; Seo, Barraclough, & Lee, 2007; Sutton & Barto, 1998), decision making should be easier and faster when this difference is larger. This predicts that RT should decrease with the absolute value of the difference in reward values on a given trial. Moreover, if reward values of objects are directly estimated/learned (object-based learning), retrieval of reward values only requires identification of the two objects. In the case of feature-based learning, however, the subject has to construct reward values of objects after retrieving their feature values and combining them. This predicts longer RT when feature-based learning is adopted, indicating that RT could reflect the dominant learning strategy on a trial-by-trial basis when both strategies are present and compete with each other to determine choice.

To test the aforementioned predictions, here we analyzed choice and RT data during a set of multi-dimensional learning and decision-making tasks that can promote different types of learning strategies.

Materials and methods

Overview of experiments and approach. Human subjects performed a set of multi-dimensional learning and decision-making tasks. During each trial, the subject selected between pairs of objects (colored shapes or patterned shapes) that would yield reward with different probabilities

(Figure 1). These probabilities could stay the same (static environment) or unpredictably change over time (dynamic environment). Moreover, the relationship between reward probabilities of objects and their features created generalizable or non-generalizable environments. In a generalizable environment, the reward probabilities assigned to different objects could be estimated from their features based on a multiplicative rule. In contrast, in a non-generalizable environment, reward probabilities assigned to different objects could not be estimated using their features. By fitting choice behavior with seven different models, we identified which of the two classes of learning strategies (feature-based or object-based) was adopted by individual subjects during each experiment. We then examined prediction of the best object-based and feature-based models on a trial-by-trial basis in order to study how RT depended on the adopted learning strategy.

Subjects. Subjects were recruited from the student population at Dartmouth College. In total, 59 subjects were recruited (34 females) to perform one or both of Experiments 1 and 2 (33 subjects performed both experiments). Data from a few subjects (8 out of 51 subjects in Experiment 1 and 19 out of 41 in Experiment 2) in which the performance was not significantly different from chance (0.5) were excluded from the analysis based on the same criteria used in our previous study (Farashahi et al., 2017). For Experiment 3, 36 additional subjects were recruited (20 females), 9 of which were excluded from the analyses because their performances were indistinguishable from chance. For Experiment 4, 36 more subjects were recruited (22 females), 11 of which were excluded from the analyses because their performances were indistinguishable from chance. No subject had a history of neurological or psychiatric illness. Subjects were compensated with a combination of money and "t-points," which are extra-credit points for classes within the Dartmouth College Psychological and Brain Sciences department. After a base rate of $10/hour or 1 t-point/hour, subjects were additionally rewarded based on their performance (e.g., number of reward points), by up to $10/hour.

Ethics statement. All experimental procedures were approved by the Dartmouth College Institutional Review Board, and informed consent was obtained from all subjects before participating in the experiment.

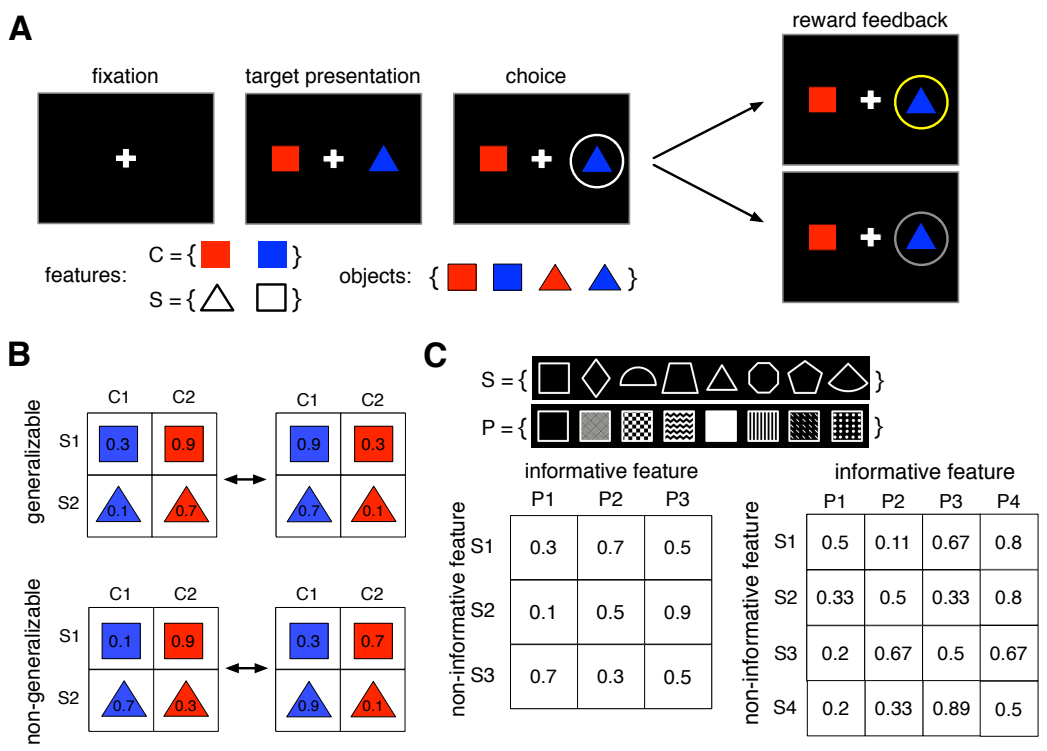

Figure 1. The schematic of the multi-dimensional learning and decision-making task. (**A**) Timeline of a trial in Experiments 1 and 2. On each trial, the subject chose between two objects (colored shapes) and was provided with reward feedback (reward or no reward) on the chosen object. The insets show the set of all features (C, color; S, shape) and objects used in Experiments 1 and 2. (**B**) Examples of the set of reward probabilities (indicated on each object) assigned to the four objects in Experiments 1 and 2. The reward probabilities assigned to the four shapes changed after every 48 trials without any cue to the subject. In the generalizable environment (Experiment 1) reward probabilities assigned to objects were predicted by feature values. In the non-generalizable environment (Experiment 2), there was no generalizable relationship between the reward values of individual objects and their features. **C.** Reward probabilities assigned to nine or sixteen possible objects during Experiments 3 and 4, respectively. Objects were defined by combinations of two features (S, shape; P, pattern), each of which could take any of three or four values in Experiments 3 and 4, respectively. The inset shows the set of shapes (top row) and patterns (bottom row) used to construct objects. Reward probabilities were assigned such that no generalizable rule predicted the reward probabilities of all objects based on feature values. In addition to choice trials similar to those in Experiments 1 and 2, Experiments 3 and 4 also included estimation trials where the subject estimated the probability of reward for an individual object by pressing one of ten keys on the keyboard.

Experiments 1 and 2. In these experiments, subjects completed two sessions (each session composed of 768 trials and lasting about one hour) of a choice task during which they selected between a pair of objects on each trial (Figure 1A). Subjects were asked to choose between objects that differed in color and shape (red triangle, blue triangle, red square, blue square) and to select the object that was more likely to provide a reward in order to maximize the total number of reward points.

During each trial, the selection of an object was rewarded independently of the reward probability of the other object. This reward schedule was fixed for a block of trials (block length, $L = 48$), after which it changed to another reward schedule without any signal to the subject. Sixteen different reward schedules were used; eight schedules consisted of generalizable rules for how combinations of features (color or shape) predicted objects' reward probabilities, and the other eight lacked generalizable rules for how combinations of features predicted objects' reward probabilities (see Farashahi et al., 2017 for more details). Therefore, the main difference between Experiments 1 and 2 was that the environments in these experiments were composed of reward schedules with generalizable and non-generalizable rules, respectively. Critically, as the subjects moved between blocks of trials, reward probabilities were reversed in order to induce volatility in reward information throughout the task. On each block of the generalizable environment (Experiment 1), feature-based rules could be used to correctly predict reward throughout the task. In the non-generalizable environment (Experiment 2), however, the reward probability assigned to each object could not be determined based on the objects' features (e.g. in a multiplicative fashion). More details about the reward schedules in Experiments 1 and 2 are provided elsewhere (Farashahi et al., 2017).

Experiment 3. In this experiment, subjects completed two sessions, each of which included 280 choice trials interleaved with five or eight short blocks of estimation trials (each block with eight trials). Subjects were asked to choose the most rewarding of two objects during each trial of the choice task. These objects were drawn from a set of eight objects constructed using combinations of three distinct patterns and three distinct shapes (Figure 1C). The two objects presented during each trial always differed in both pattern and shape. Other aspects of the choice task were similar to those in Experiments 1 and 2, except that reward feedback was given for both objects rather than just the chosen object in order to accelerate learning. Subjects then provided estimates of

each object's reward during estimation trials. Possible values for these estimates ranged from 5% to 95% in 10% increments. All subjects completed five blocks of estimation trials throughout the task (after trials 42, 84, 140, 210, and 280 of the choice task), and some subjects had three additional blocks of estimation trials (after trials 21, 63, and 252) to better assess changes in estimations over time. Each session of the experiment was approximately 45 minutes in length, with a break between sessions. The second session was similar to the first, but with different sets of shapes and patterns.

Selection of a given object was rewarded (independently of the other object on a given trial) based on a reward schedule with a moderate level of generalizability such that only some objects' reward probabilities could be determined based on their feature values. In Experiment 3, one feature (shape or pattern) was informative about reward probability while the other feature was not. Although the informative feature was predictive of reward on average, this prediction was not generalizable. Additionally, the non-informative feature did not follow any generalizable rule (i.e. average reward probability across objects with similar feature value was equal to 0.5) (Figure 1C). This reward schedule ensured that subjects could not use a generalizable feature-based rule to accurately predict reward probability for all objects. Similarly to Experiments 1 and 2, the informative feature was randomly assigned and counter-balanced across subjects to minimize the effects of intrinsic pattern or shape biases. More details about the reward schedules are provided elsewhere (Farashahi et al., 2017).

Experiment 4. This experiment was similar to Experiment 3, except that each feature could take on one of four feature values for each feature (i.e. there were four shapes and four patterns), resulting in an environment with higher dimensionality. Each subject completed two sessions of 336 choice trials interleaved with five or eight short blocks of estimation trials (each block with eight trials). The objects in this experiment were drawn from a set of twelve objects, which were combinations of four distinct patterns and four distinct shapes. The probabilities of reward for different objects (reward matrix) were set such that there was one informative feature and one non-informative feature, just as in Experiment 3 (Figure 1C). Minimum and maximum average reward values for features were similar for Experiments 3 and 4.

Model fitting. To capture subjects' learning and choice behavior, we used seven different reinforcement learning (RL) models based on object-based or feature-based learning, which rely on different assumptions about how reward values are represented and updated. These models were fit to experimental data by minimizing the negative log likelihood of the predicted choice probability given different model parameters using the 'fminsearch' function in MATLAB (MathWorks, Inc., Natick, MA). Three measures of goodness-of-fit were used to determine the best model to account for the behavior in each experiment: average negative log likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC). The details of these models are provided elsewhere (Farashahi et al., 2017).

Analysis of response time. For most analyses, response time (RT) was normalized (z-scored) by first subtracting the mean RT of a given subject from their RT values and then dividing the outcome by the standard deviation of RT for that subject. This normalization allowed us to directly compare data from all subjects. We removed trials with RT less than 50 msec and more than 5 sec since they correspond to choices with no deliberation and when the subject was distracted, respectively. We used several generalized linear models (GLM) with the following regressors (all z-scored) to predict the normalized RT on each trial: 1) the difference between the actual or estimated (using the model that provide the best fit of choice data) reward probabilities of the two objects presented on a given trial; 2) the trial number within a block; 3) reward outcome on the preceding trial; and 4) a 'model-adoption' index comparing the prediction of the best feature-based and object-based models on a given trial. The model-adoption index was defined as the difference between the goodness-of-fit by the best feature-based and object-based models based on the BIC per trial (= -2*Log likelihood $+2k*\log(N_{trials})/N_{trials}$,where k indicates the number of parameters in a given model). The best object-based and feature-based models were determined based on the BIC for each subject in a given experiment. We note that even though we used BIC for these analyses, similar results were obtained with AIC. Thus the model-adoption index determines the degree to which feature-based or object-based models predicted choice behavior in a given trial more accurately for an individual subject (positive values correspond to a better prediction by the object-based model).

To further investigate the relationship between RT and the adopted model on a given trial, we grouped trials into object-based, feature-based, and equivocal trials depending on which model

provided a better fit (based on BIC) and whether the absolute value of model-adoption index was larger than a threshold. More specifically, equivocal trials were defined as trials in which the absolute value of the model-adoption index was less than 0.1. This resulted in assigning approximately 39, 39, and 22 percent of trials as object-based, feature-based, and equivocal, respectively (over all experiments).

Results

To examine how RT depends on the construction and comparison of reward values, we used various reinforcement learning models to determine whether object-based or feature-based learning was adopted by individual subjects in a given experiment. For each model, we computed three measures of goodness-of-fit in order to determine the best overall model to account for the behavior in each experiment (Table 1). The smaller value for each measure indicates a better fit of choice behavior.

This analysis showed that in Experiment 1 (dynamic generalizable environment), the feature-based models provided better fits than the object-based models, indicating that subjects adopted feature-based learning more often in this experiment. In contrast, subjects more frequently adopted object-based learning in Experiment 2 (dynamic non-generalizable environment). In Experiment 3 (static and non-generalizable environment), subjects more frequently adopted object-based learning. An examination of fits and the pattern of estimations over time revealed that in this experiment, subjects adopted feature-based learning first before slowly switching to object-based learning (Farashahi et al., 2017). In contrast, in Experiment 4 (static non-generalizable environment), which involved a larger number of objects and slightly larger generalizability than Experiment 3, subjects more frequently adopted feature-based learning and did not switch to object-based learning.

Altogether, the analysis of subjects' choice behavior demonstrates that subjects adopted different learning strategies depending on the structure of reward in each environment. As mentioned above, each of these learning strategies requires a different method for the construction and/or comparison of reward value. Object-based learning relies on directly representing and updating

the rewards of individual options. In contrast, feature-based learning requires representing and updating the average reward values of all features, and then combining these values to estimate the reward values for individual options. To test whether learning strategy influenced RT, we next examined the pattern of RT in our four experiments.

Model	Coupled feature-based	Uncoupled feature-based	Feature-based with decay	Coupled object-based	Uncoupled object-based	Object-based with decay	
# pars.	5	5	6	4	4	5	
-LL	435.3±11.6***	458.6±8.9***	434.5±11***	451.8±9	474.1±7.4	450.7±8.7	Exp. 1
AIC	880.5±23.1***	927.1±17.8***	880.9±21.9***	911.5±17.9	956.1±14.8	911.4±17.4	
BIC	903.7±23.1*	950.3±17.8***	908.8±21.9**	930.1±17.9	974.7±14.8	934.6±17.4	
-LL	478.2±7.7	493.2±5.3	477.2±7.6	466.4±7.8	494.8±5.5	462.4±7.8*	Exp. 2
AIC	966.5±15.4	996.5±10.6	966.4±15.2	940.8±15.7^{+}	997.6±10.9	934.7±15.6*	
BIC	989.7±15.4	1019.7±10.6	994.3±15.2	959.3±15.7^{+}	1016.1±10.9	957.9±15.6*	
-LL	328.1±4.8	337.0±5.4	323.3±5.6	330.0±6.9	331.6±5.6	290.8±7.9**	Exp. 3
AIC	666.2±9.6	683.9±10.9	658.7±11.2	668.0±13.9	671.2±11.3	591.7±15.7**	
BIC	687.9±9.6	705.6±10.9	684.6±11.2	685.3±13.9	688.5±11.3	613.3±15.7***	
-LL	377.6±6.5***	378.5±6.8***	336.6±9.1**	409.3±4.1	409.6±3.9	350.9±6.8	Exp. 4
AIC	765.2±13.1***	767.1±13.7***	685.2±18.1**	826.6±8.3	827.2±7.8	711.9±13.7	
BIC	787.8±13.1***	789.6±13.7**	712.3±18.1*	844.6±8.3	845.3±7.8	734.4±13.7	

Table 1. The average values of three measures of goodness-of-fit (-log likelihood (-LL), Akaike information criterion (AIC), and Bayesian information criterion (BIC)) based on three feature-based RLs and their object-based counterparts, separately for each of four experiments. The average values are computed over all subjects (mean ± s.e.m). The best model (feature-based RL with decay) and its object-based counterpart are highlighted in cyan and brown, respectively. Each feature-based RL was compared with its object-based counterpart using a two-tailed, sign-rank test. The significance level of the test is coded as: $0.05 < p < 0.1$ ($^{+}$), $0.01 < p < 0.05$ (*), $0.001 < p < 0.01$ (**), and $p < 0.001$ (***).

The average RTs across all subjects (using median RT of all trials for each subject) were equal to 0.78±0.04s, 0.85±0.05s, 1.04±0.05s, and 1.01±0.03s for Experiments 1 to 4, respectively. A longer RT for Experiments 3 and 4 could be due to a larger number of objects in those experiments (8 and 12, respectively, in Experiments 3 and 4, rather than 4 in Experiments 1 and 2). Experiments 1 and 2 were identical except for the way in which reward probabilities assigned to the four objects resulting in generalizable and non-generalizable environments. Because of the non-generalizable reward schedules in Experiment 2, this experiment was more difficult to perform than Experiment 1. This was reflected in a lower average performance (harvested reward) (Wilcoxon rank-sum test, $p < 0.001$; the average performance for Experiments 1 and 2 were equal to 0.564±0.004 and 0.536±0.003, respectively) and a larger number of subjects whose performance did not differ from chance in Experiment 2 relative to Experiment 1 (Pearson's chi-square test, $P = .0014$; 8 of 51 in Experiment 1, and 19 of 41 in Experiment 2).

Despite the fact that Experiment 2 was more difficult, the average RT over all subjects in Experiment 2 was not statistically different from that in Experiment 1 (Wilcoxon rank-sum test, $P = 0.2$). This observation can be explained by noting that subjects adopted object-based learning in Experiment 2 more often than in Experiment 1 (74% vs. 38% in Experiments 2 and 1, respectively), and the average RT for subjects who adopted object-based learning was smaller than that for those who adopted feature-based learning in both experiments (two-sided Wilcoxon rank-sum test, Experiment 1: $P = 0.01$, Experiment 2: $P = 0.04$). More specifically, the average median RTs for subjects who adopted object-based and feature-based learning in Experiment 1 were equal to 0.71±0.06s and 0.85±0.05s, respectively. The average median RTs for subjects who adopted object-based and feature-based learning in Experiment 2 were equal to 0.79±0.07s and 0.88±0.07s, respectively. These results provide preliminary evidence supporting our hypothesis that construction and comparison of reward value is faster when the object-based strategy is adopted.

Given these results, we next utilized generalized linear models (GLMs) to examine the influence of several factors on the pattern of RT in each experiment. These included the absolute value of the difference in the actual or estimated probability of reward on the two objects presented on a given trial (absolute difference in reward probability), the trial number within a given experiment, reward outcome on the preceding trial, and the model-adoption index (see Methods).

We used normalized RT (z-scored for each subject) in order to analyze data from all subjects together.

A

Regressor	abs. difference in actual reward prob.	trial number	BIC(Ft) – BIC(Obj)	reward outcome on prev. trial	R^2
Exp. 1	-0.05±0.005***	-0.15±0.006***	0.02±0.005***	-0.03±0.006***	0.027
Exp. 2	-0.07±0.008***	-0.11±0.008***	0.02±0.008**	-0.04±0.008***	0.020
Exp. 3	-0.06±-0.008***	-0.16±0.008***	0.03±0.008**	-0.05±0.008***	0.031
Exp. 4	-0.04±-0.007***	-0.23±0.007***	0.01±0.007*	-0.03±0.007***	0.057

B

Regressor	abs. difference in subjective reward prob.	trial number	BIC(Ft) – BIC(Obj)	reward outcome on prev. trial	R^2
Exp. 1	-0.13±0.006***	-0.16±0.005***	0.02±0.005***	-0.05±0.006***	0.041
Exp. 2	-0.13±0.008***	-0.12±0.008***	0.03±0.008***	-0.07±0.008***	0.031
Exp. 3	-0.21±-0.008***	-0.14±0.008***	0.06±0.008**	-0.04±0.008***	0.073
Exp. 4	-0.18±-0.007***	-0.21±0.007***	0.01±0.007*	-0.03±0.007***	0.089

Table 2. GLMs reveal different factors influencing RT. We used various GLMs to predict the normalized RT as a function of the absolute difference in the actual (A) and estimated (B) probability of reward on the two objects presented on a given trial (abs. difference in reward prob.), the trial number within a block of the experiment, the difference between BIC per trial based on the best feature-based and object-based models (i.e., model-adoption index) for a given subject, and reward outcome on the preceding trial. Reported values are the normalized regression coefficients (±s.e.m.) and the significance level of the two-tailed sign-rank test is coded as: $0.01 < p < 0.05$ (*), $0.001 < p < 0.01$ (**), and $p < 0.001$ (***). Interaction terms which were not statistically significant are not reported.

The results of the GLM fits revealed several important aspects of RT in our experiments (Table 2). First, subjects became faster over time in all experiments by an average of 300-400ms at the end relative to the beginning of the experiments (Figure 2). Moreover, in Experiments 1 and 2, in which there was a reversal in reward probabilities after every block of 48 trials (dynamic environments), RT increased at the beginning of each block and then decreased as subjects

learned the reward probabilities associated with the four objects in each block (Figure 3). This resulted in average of 20ms and 40ms decrease in RT over the course of each block in Experiments 1 and 2, respectively.

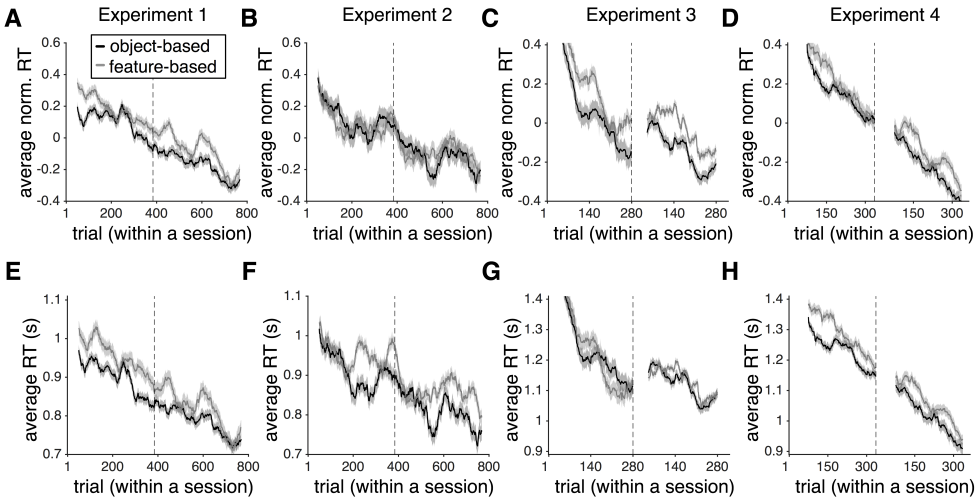

Figure 2. Subjects made decisions more quickly as they progressed through the experiments, and during trials when they used the object-based rather than the feature-based model. (**A-D**) Plotted is the average z-scored response time (across all subjects) as a function of the trial number within a session of an experiment separately for trials in which the object-based or feature-based model was used. Panels A-D correspond to Experiments 1 to 4, respectively, and shaded areas indicate s.e.m. The gray dotted line shows the time point during a session (in terms of trial number) at which the experiment was paused for a short rest. In Experiments 3 and 4, a different set of objects was introduced at this point. (**E-H**) The same as in A-D but showing the average actual RT across all subjects. Overall, subjects made decisions 300-400ms faster at the end relative to the beginning of the experiments. Nevertheless, they were consistently faster on trials when the adopted object-based than feature-based strategy.

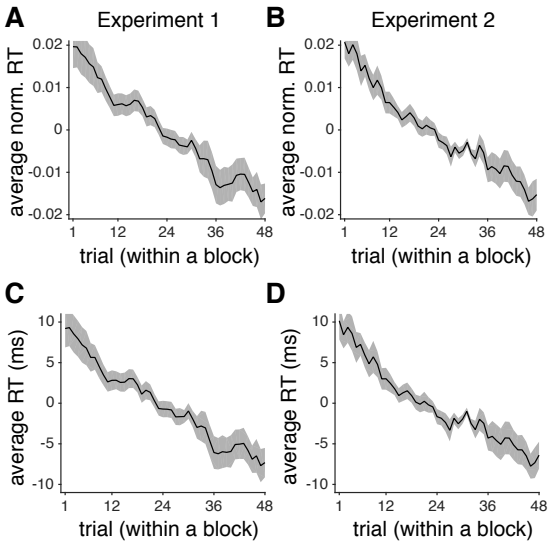

Figure 3. Subjects made decisions more quickly as they progressed through the trials within each block of Experiments 1 and 2. Plotted is the average z-scored response time (across all subjects) as a function of the trial number within a session of experiment 1 (**A**) and experiment 2 (**B**). Shaded areas indicate s.e.m. (**C-D**) The same as in A-B but showing the average actual RT over all subjects.

Second, we found that the absolute difference in the actual or estimated reward probabilities negatively affected RT such that RT was maximal when this difference was close to zero (Table 2; Figures 4, 5). Importantly, the absolute difference in estimated reward probabilities (based on the fit of choice behavior of each subject) was a better predictor than the absolute difference in the actual reward probabilities (compare R^2 based on the GLMs in panels A and B of Table 2). The absolute difference in the estimated reward probabilities resulted in about 120ms change in RT (Figure 5). Third, the reward outcome on the previous trial also had a negative effect on RT such that RT was shorter after rewarded trials compared to after unrewarded trials. Overall, RT on trials following rewarded trials was shorter than RT following unrewarded trials by 33±12ms, 50±24ms, 50±11ms, and 16±13ms in Experiments 1 to 4, respectively, indicating that subjects spent more time making a choice following unrewarded trials.

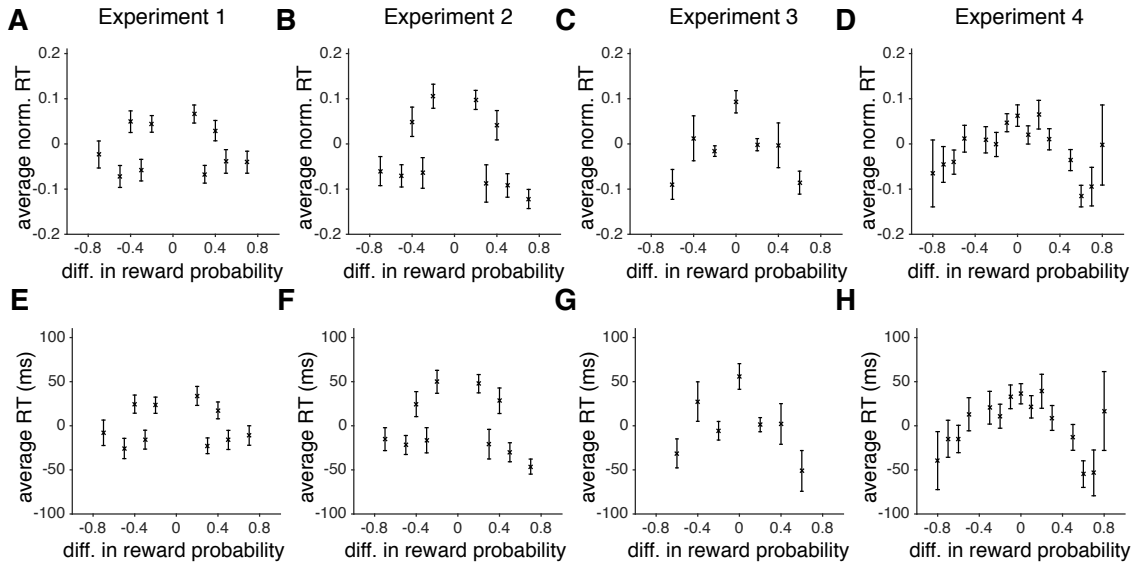

Figure 4. Subjects were faster on trials in which the difference between the actual reward probabilities of the competing objects deviated more from zero. (**A-D**) Panels A-D correspond to Experiments 1 to 4, respectively, and the error bars indicate s.e.m. (**E-H**) The same as in A-D but showing the average actual RT (median removed) over all subjects. Overall, the difference in actual reward probabilities of the two alternative objects resulted in up to ~100 ms increase in RT when reward probabilities were close to each other relative to when they were very different.

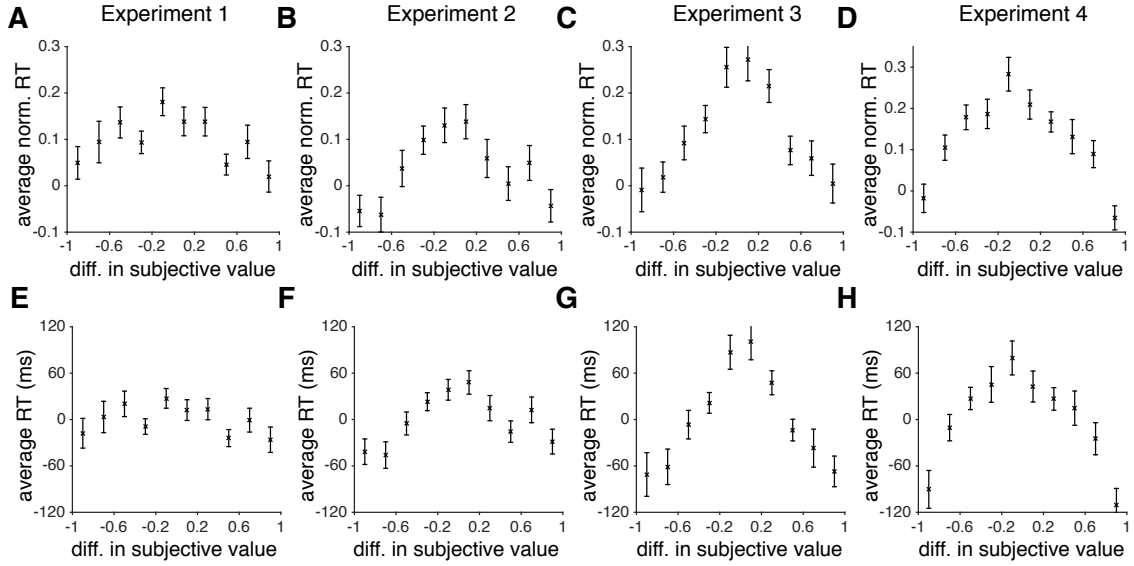

Figure 5. Subjects were faster on trials in which the difference between the estimated reward probabilities of the competing objects deviated more from zero. (**A-D**) Panels A-D correspond to Experiments 1 to 4, respectively, and the error bars indicate s.e.m. (**E-H**) The same as in A-D but showing the average actual RT (median removed) over all subjects. Overall, the difference in estimated reward probabilities of the two alternative objects resulted in up to ~150 ms increase in RT when reward probabilities were close to each other relative to when they were very different.

Finally, the model-adoption index, which indicates the degree to which a learning strategy (object-based or feature-based) was adopted by an individual subject on a given trial, negatively influenced RT (Table 2). This illustrates that subjects were slower on trials in which the feature-based model provided a better fit (i.e., when they adopted feature-based strategy). To further investigate this relationship, we computed RT for object-based, feature-based, and equivocal trials (trials where the adopted model was ambiguous; see Methods). This analysis illustrated that subjects were slower by about 40ms when they adopted a feature-based compared to an object-based strategy (Figure 6A-B; Table 3). Moreover, RT was fastest on equivocal trials. At the same time, however, the minimum BIC based on the two best models on each trial, a measure of the best prediction by either model, was the largest for equivocal trials (Figure 6C; Table 3). These indicate that subjects were slightly faster on equivocal trials, perhaps because on those trials they made decisions more randomly as evident from reduced quality of fit. Overall, these results illustrate that RT is influenced by the learning strategy adopted by subjects on a trial-by-trial basis.

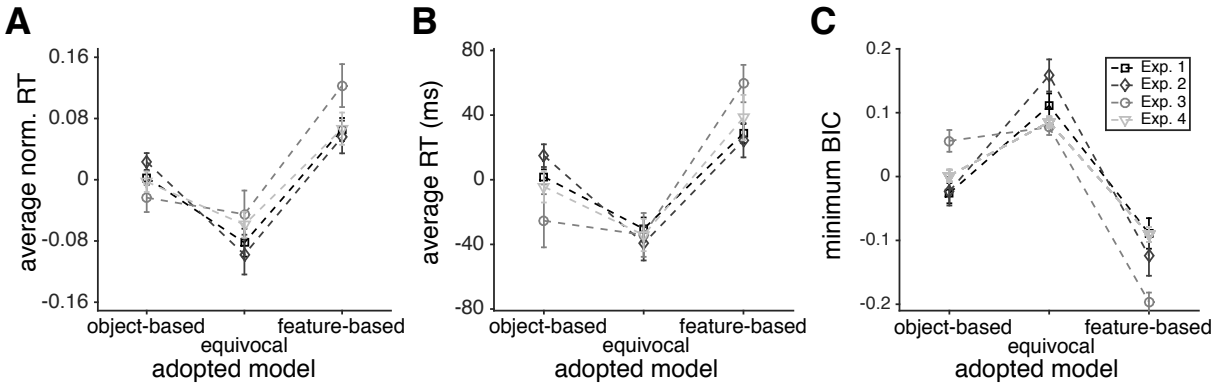

Figure 6. The response time depended on the model adopted by individual subjects on a trial-by-trial basis. (**A**) Plotted is the average (across subjects) normalized RT for three types of trials. Subjects made decisions faster during trials where the object-based model was used than when the feature-based model was used. In some experiments, subjects were even faster on equivocal trials (when the better model was ambiguous). (**B**) The same as in A shows the average actual RT over all subjects after removing the median RT from each individual RT distribution. **(C)** Plotted is the average (across subjects) of minimum BIC for each group of trials after removing the average minimum BIC over all trials for each subject. Subjects' choice behavior was hardest to fit on equivocal trials indicating that they made choices more stochastically on those trials.

	Object-based vs. equivocal	Object-based vs. Feature-based	Feature-based vs. equivocal	
Exp. 1	0.0023	0.0039	0.0001	average RT (ms)
Exp. 2	0.0024	0.2305	0.0051	
Exp. 3	0.6139	0.0046	0.0003	
Exp. 4	0.0578	0.0480	0.0009	
Exp. 1	0.0045	0.0030	0.0001	average norm. RT
Exp. 2	0.0026	0.2443	0.0064	
Exp. 3	0.5806	0.0062	0.0003	
Exp. 4	0.0653	0.0544	0.0010	
Exp. 1	0.0036	0.0001	0.1231	minimum BIC
Exp. 2	0.0001	0.0001	0.0765	
Exp. 3	0.5615	0.0001	0.0001	
Exp. 4	0.0002	0.0001	0.0032	

Table 3. Comparisons of the average RT and minimum BIC during different types of trials. Reported are the *P*-values for comparisons of the average RT, average normalized RT, and minimum BIC between a given pair of trials (object-based, feature-based, and equivocal trials) depicted in Figure 6.

Because subjects also became faster over time, a shorter RT on trials when the object-based strategy was adopted could be attributed to a gradual shift from using feature-based to object-based strategy (especially in Experiment 3 in which this shift was prominent). To test this possibility, we computed the average RT over the course of experiment across all subjects separately for trials in which object-based or feature-based strategy was adopted (Figure 2). We found a consistently larger RT on feature-based trials during the entire course of each experiment. This indicates that the overall faster RT on object-based trials was not caused by a combination of a shift to object-based learning and faster decision making as subjects progressed through each experiment.

Discussion

We investigated the pattern of RT during a set of a multi-dimensional learning and decision-making tasks in order to explore neural processes underlying the construction and comparison of reward values. Specifically, we tested whether the type of learning strategies adopted by the subjects and the underlying reward value representation were reflected in the pattern of RT during binary choice. We found a decrease in RT over the course of experiments and after rewarded trials. Moreover, RT was strongly modulated by the difference in the value of reward probabilities assigned to the two alternative objects on a given trial such that it was maximal when the absolute value of the difference in the actual or estimated reward probabilities was minimal. As expected, this effect was stronger for estimated than for actual reward probabilities. More importantly, the model adopted by the subject on a trial-by-trial basis influenced RT such that subjects were faster when they made a choice using the object-based rather than the feature-based strategy. These results confirm our hypotheses that at least two stages of value-based decision making, construction of reward value for each option, and value comparison between pairs of options are influenced by the learning strategy adopted by the subject. A crucial aspect of our experiments, which allowed us to study the influence of value construction on RT, is the rich multidimensional task in which multiple learning strategies coexist and interact to determine choice.

Our results are relevant to visual search and object identification based on reward values (Wolfe, 2013). We found that RT for selection between two options increased as the subjective values of the two options became more similar. If we consider selection between the two objects on a given trial as a search for a more rewarding object (target), our observation dovetails with the finding in visual search that more visual similarity between the target and distractor increases RT (Duncan & Humphreys, 1989). Interestingly, there are a number of studies in which the visual search has been manipulated by differential reward assignment on targets or distractors (Anderson, Laurent, & Yantis, 2011; Hickey, Chelazzi, & Theeuwes, 2010; Hickey & Peelen, 2015; Theeuwes & Belopolsky, 2012). Overall, those studies have demonstrated that task-irrelevant distractors previously associated with larger rewards can slow down target detection (Anderson et al., 2011; Hickey & Peelen, 2015), and this reduction in performance has been attributed to attentional capture by the rewarding feature. Interestingly, in an experiment in

which different reward magnitudes (large and small) were randomly delivered on correct trials, the amount of reward differentially modulated RT on the following trial depending on whether the rewarding feature of the target stayed the same or not between the two trials (Hickey et al., 2010). If the feature stayed the same, subjects were faster after high-reward trials relative to low-reward trials. In contrast, if the feature switched, subjects were slower after high-reward trials relative to low-reward trials.

The feature-based learning provides a heuristic for learning and choice in dynamic, multi-dimensional environments. Importantly, heuristic feature-based learning not only reduces dimensionality but also increases adaptability without compromising precision (Farashahi et al., 2017). Here, we showed that subjects were faster when they made a choice using the object-based rather than the feature-based model. Based on these results, we predict that there will be a larger difference in RT when feature-based versus object-based strategy is adopted for more high-dimensional options. However, RT using the feature-based strategy could be dramatically reduced for high-dimensional options if the comparison between the two options is limited to dissimilar features only, allowing the feature-based learning to be a "fast and frugal" heuristic (Gigerenzer and Goldstein, 1996). The dependence of RT on the subject's adopted model of the environment also implies that the adopted model should be estimated and considered when interpreting the top-down effects during conventional or foraging visual search tasks (Wolfe, 2013; see Khorsand, Moore, & Soltani, 2015 for a review of top-down effects).

Finally, with regard to how value comparison contributes to RT during decision making, we found that the difference in the probability of reward assigned to the two targets presented on each trial predicted RT, similarly to our previous observation (Soltani, De Martino, & Camerer, 2012). There is growing interest in economics to use RT in order to infer a decision maker's preference beyond what can be revealed only by choice (Clithero, 2016a, 2016b; Konovalov & Krajbich, 2016; Spiliopoulos & Ortmann, 2016). For example, based on the observation that RT increases as the subjective values of options become closer to each other in a binary choice, Konovalov and Krajbich (Konovalov & Krajbich, 2016) have suggested that RT could be used to detect the indifference point. In another work, Clithero (Clithero, 2016a) has suggested that RT can be used along with choice to improve out-of-sample predictions. Our results suggest that even the adopted model could also be predicted based on RT on a given trial. It would be

interesting to test in the future whether adding RT can improve the fit of choice behavior. However, as we showed here, RT monotonically decreases over time even during continuous learning, indicating that precautions must be taken if RT is used as a measure of the indifference point in such situations.

Altogether, using behavior during a set of multi-dimensional learning and decision-making tasks, we show that the pattern of RT can be used to investigate different stages of value-based leaning and decision making, and more specifically how reward value is constructed and compared on a single trial.

References

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367–10371.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410.

Breitmeyer, B. G. (1975). Simple reaction time as a measure of the temporal response properties of transient and sustained channels. *Vision Research*, *15*(12), 1411–1412.

Clithero, J. A. (2016a). Improving out-of-sample predictions using response times and a model of the decision process. *Working Paper. Available at SSRN: Https://Ssrn.com/abstract=2798459*.

Clithero, J. A. (2016b). Response times in economics: Looking through the lens of sequential sampling models. *Working Paper. Available at SSRN: Https://Ssrn.com/abstract=2795871*.

Corrado, G. S., Sugrue, L. P., Seung, H. S., & Newsome, W. T. (2005). Linear-Nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior*, *84*(3), 581–617.

Donders, F. C. (1868). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision *Nature Communications* 8:1768.

Fellows, L. K. (2006). Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain*, *129*(4), 944–952.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Hickey, C., Chelazzi, L., & Theeuwes, J. (2010). Reward changes salience in human vision via the anterior cingulate. *Journal of Neuroscience*, *30*(33), 11096–11103.

Hickey, C., & Peelen, M. V. (2015). Neural mechanisms of incentive salience in naturalistic human vision. *Neuron*, *85*(3), 512–518.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., … Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron*, *90*(1), 177–190.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a'preattentive' feature search task. *Nature*, *387*(6635), 805–807.

Khorsand, P., Moore, T., & Soltani, A. (2015). Combined contributions of feedforward and feedback inputs to bottom-up attention. *Frontiers in Psychology*, 86.

Konovalov, A., & Krajbich, I. (2016). Revealed Indifference: Using Response Times to Infer Preferences. *Working Paper. Columbus, OH: The Ohio State University.*

Krummenacher, J., Grubert, A., & Müller, H. J. (2010). Inter-trial and redundant-signals effects in visual search and discrimination tasks: Separable pre-attentive and post-selective effects. *Vision Research*, *50*(14), 1382–1395.

Krummenacher, J., Müller, H. J., & Heller, D. (2001). Visual search for dimensionally redundant pop-out targets: Evidence for parallel-coactive processing of dimensions. *Perception & Psychophysics*, *63*(5), 901–917.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308.

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: A model and a method. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 276–291.

Louie, K., & Glimcher, P. W. (2010). Separating value from choice: Delay discounting activity in the lateral intraparietal area. *Journal of Neuroscience*, *30*(16), 5498–5507.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

Seo, H., Barraclough, D. J., & Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex*, *17*, 110–117.

Soltani, A., De Martino, B., & Camerer, C. (2012). A range-normalization model of context-dependent choice: a new model and evidence. *PLoS Comput Biol*, *8*(7), e1002607.

Soltani, A., & Wang, X.-J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *The Journal of Neuroscience*, *26*(14), 3731–3744.

Soltani, A., & Wang, X.-J. (2008). From biophysics to cognition: reward-dependent adaptive choice behavior. *Current Opinion in Neurobiology*, *18*, 209–216.

Spiliopoulos, L., & Ortmann, A. (2016). The BCD of response time analysis in experimental economics. *Working Paper. Available at SSRN: Https://Ssrn.com/abstract=2401325*.

Sternberg, S. (1969a). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Sternberg, S. (1969b). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies of valuation and decision making, *6*, 363–375.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tanaka, Y., & Shimojo, S. (1996). Location vs feature: Reaction time reveals dissociation between two visual functions. *Vision Research*, *36*(14), 2125–2140.

Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: oculomotor capture by rewarding stimuli. *Vision Research*, *74*, 80–85.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 459–478.

Tynan, P. D., & Sekuler, R. (1982). Motion processing in peripheral vision: Reaction time and perceived velocity. *Vision Research*, *22*(1), 61–68.

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition*, *33*(2), 213–220.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, *13*(3), 1–17.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419–433.