Spring 2022

Testing Models of Context-Dependent Outcome Encoding in Reinforcement Learning

William M. Hayes IV

Testing Models of Context-Dependent Outcome Encoding in Reinforcement Learning

by

William M. Hayes IV

Bachelor of Science
North Greenville University, 2014

Master of Arts
University of North Carolina Wilmington, 2017

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Experimental Psychology

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

Douglas H. Wedell, Major Professor

Svetlana V. Shinkareva, Committee Member

Amit Almor, Committee Member

Stefano Palminteri, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

There are several people who supported and assisted me through the process of writing this dissertation. First, I would like to thank my doctoral adviser, Dr. Douglas H. Wedell, for mentoring me over the last five years. I am extremely grateful to Doug for sparking my interest in cognitive modeling and for encouraging me to pursue this line of research. Being able to combine his expertise in range-frequency theory and context effects with my interest in reinforcement learning for this project has been the highlight of my graduate school career. Thank you, Doug, for being an outstanding mentor.

I would also like to thank the members of my dissertation committee. Drs. Svetlana V. Shinkareva, Amit Almor, and Stefano Palminteri provided feedback and encouragement that was very helpful to me as I worked on this project. I truly appreciate their insightful questions, edits, and suggestions. Special thanks to Dr. Palminteri, who agreed to serve on my committee despite the 6-hour time difference between Columbia, SC and Paris.

Finally, I would like to thank my family and loved ones who supported and encouraged me through this process. Thank you, Mom and Dad, for always believing in me, for teaching me the value of hard work, and for pushing me to be the best that I can be at whatever I set out to do. This would not have been possible without you.

ABSTRACT

Previous studies of reinforcement learning (RL) have established that choice outcomes are encoded in a context-dependent fashion. Several computational models have been proposed to explain context-dependent encoding, including reference point centering and range adaptation models. The former assumes that outcomes are centered around a running estimate of the average reward in each choice context, while the latter assumes that outcomes are compared to the minimum reward and then scaled by an estimate of the range of outcomes in each choice context. However, there are other computational mechanisms that can explain context dependence in RL. In the present study, a frequency encoding model is introduced that assumes outcomes are evaluated based on their proportional rank within a sample of recently experienced outcomes from the local context. A hybrid range-frequency model is also considered that combines the range adaptation and frequency encoding mechanisms. We conducted two fully incentivized behavioral experiments using choice tasks for which the candidate models make divergent predictions. The results were most consistent with models that incorporate frequency or rank-based encoding. The findings from these experiments deepen our understanding of the underlying computational processes mediating context-dependent outcome encoding in human RL.

TABLE OF CONTENTS

Acknowledgements iii

Abstract iv

List of Tables vi

List of Figures vii

Chapter 1: Introduction 1

1.1 Model Descriptions 3

1.2 Distinguishing among Context-Dependent Encoding Models 16

Chapter 2: Experiment 1 21

2.1 Model Predictions 24

2.2 Method 32

2.3 Results 39

2.4 Discussion 50

Chapter 3: Experiment 2 53

3.1 Model Predictions 55

3.2 Method 62

3.3 Results 69

3.4 Discussion 83

Chapter 4: General Discussion 86

References 97

Appendix A: Additional Models 106

LIST OF TABLES

Table 1.1 Summary of an Instrumental Learning Task from a Prior Study with Model Predictions 17

Table 2.1 Summary of the Instrumental Learning Task in Experiment 1 23

Table 2.2 Model Predictions for the Target Choice Pairs in Experiment 1 32

Table 2.3 Model Comparison Results in Experiment 1 43

Table 2.4 Mean Parameter Estimates in Experiment 1 44

Table 3.1 Summary of the Instrumental Learning Task in Experiment 2 54

Table 3.2 Model Comparison Results in Experiment 2 75

Table 3.3 Mean Parameter Estimates in Experiment 2 76

Table A.1 Additional Model Comparison Results 107

LIST OF FIGURES

Figure 2.1 REFERENCE Model Simulations for Experiment 1 26

Figure 2.2 RANGE Model Simulations for Experiment 1..... 27

Figure 2.3 FREQUENCY Model Simulations for Experiment 1..... 29

Figure 2.4 RANGE-FREQUENCY Model Simulations for Experiment 1..... 31

Figure 2.5 Trial Timeline for Experiment 1 35

Figure 2.6 Learning and Transfer Phase Results for Experiment 1 40

Figure 2.7 Transfer Phase Pairwise Choice Preferences
in Experiment 1: Empirical Data vs. Model Simulations 45

Figure 2.8 Responses to the Post-Task Questions in Experiment 1..... 48

Figure 3.1 REFERENCE Model Simulations for Experiment 2 57

Figure 3.2 RANGE Model Simulations for Experiment 2..... 58

Figure 3.3 FREQUENCY Model Simulations for Experiment 2..... 60

Figure 3.4 RANGE-FREQUENCY Model Simulations for Experiment 2..... 62

Figure 3.5 Trial Timeline for Experiment 2..... 66

Figure 3.6 Learning and Transfer Phase Results for Experiment 2 71

Figure 3.7 Learning Phase Choice Behavior in Experiment 2:
Empirical Data vs. Model Simulations..... 77

Figure 3.8 Transfer Phase Pairwise Choice Preferences
in Experiment 2: Empirical Data vs. Model Simulations 79

Figure 3.9 Transfer Phase Choice Rates in Experiment 2:
Empirical Data vs. Model Simulations..... 80

Figure 3.10 Responses to the Post-Task Questions in Experiment 2..... 82

Figure A.1 Additional Model Simulations in Experiment 1.....108

Figure A.2 Additional Model Simulations in Experiment 2: Learning Phase.....109

Figure A.3 Additional Model Simulations in Experiment 2: Transfer Patterns ..110

Figure A.4 Additional Model Simulations in Experiment 2: Choice Rates 111

CHAPTER 1

INTRODUCTION

Normative theories of value-based decision-making assume that reward values are encoded in a context-independent fashion (Luce, 1959; von Neumann & Morgenstern, 1944). This means that the cognitive representation of a fixed reward should not depend on the values of other rewards in its environment. Context independence would seem to be a prerequisite for making rational, reward-maximizing decisions. However, there is ample behavioral and neuroscientific evidence for context-dependent valuation across a variety of species, including humans, primates, birds, and insects (e.g., Burke et al., 2016; Mullet & Tunney, 2013; Padoa-Schioppa, 2009; Palminteri et al., 2015; Pompilio & Kacelnik, 2010; Shafir et al., 2002; Tobler, Fiorillo, & Schultz, 2005; Tremblay & Schultz, 1999). These studies have shown that rewards are represented in a way that critically depends on other rewards in the choice environment, so that the same reward will be evaluated differently across different contexts.

Why might the brain encode values on a context-dependent scale? Consider that the range of values that organisms might encounter in their environment is theoretically infinite, yet neurons have a finite range of firing rates for encoding these values. This biological constraint is problematic for theories that assume absolute value representations (Mullett & Tunney, 2013). A more efficient neural code could be implemented by adapting firing rates to the

distribution of rewards in the local context (Louie & Glimcher, 2012; Seymour & McClure, 2008). Although it can increase sensitivity to value differences within contexts, context dependence can also result in suboptimal choice behavior when values are extrapolated outside of their original encoding contexts (Bavard et al., 2018; Klein et al., 2017; Palminteri et al., 2015). However, there is evidence that the human brain encodes a combination of absolute and relative values, and that it can flexibly—and perhaps rationally—switch between coding schemes depending on features of the choice environment and task demands (Burke et al., 2016; Pischedda et al., 2020; for behavioral evidence, see Juechems et al., 2021).

The present study focuses on context-dependent value encoding in reinforcement learning (RL). RL is the process through which organisms learn to predict the consequences of their actions and adjust their choice behavior to maximize rewards and minimize punishments (Dayan & Niv, 2008). Previous RL studies have shown that when choice options are encountered repeatedly in separate contexts, people learn the values of those options in a context-dependent fashion (Bavard et al., 2018; 2021; Hayes & Wedell, in press; Klein et al., 2017; Palminteri et al., 2015). Context-induced biases are revealed when the options are transferred out of their original learning contexts and reencountered in novel contexts. Several RL models have been proposed to explain these effects, and the most prominent have relied on two different computational mechanisms: reference point centering (Palminteri et al., 2015) and range adaptation (Bavard et al., 2018; 2021). While these models offer a parsimonious

account of the context effects observed in prior studies (for a review, see Palminteri & Lebreton, 2021), there are other potential mechanisms that could produce the observed behavior but have not received as much attention in the literature. The purpose of the present study is to test a larger set of models using choice tasks that can dissociate several context-dependent mechanisms simultaneously. The following section introduces the models that will be considered.

1.1 Model Descriptions

RL models describe how decision-making agents learn from previous choice outcomes to select options that maximize expected rewards (Rangel et al., 2008; Sutton & Barto, 1998). In the present study, choice options are encountered in separate groupings, or contexts, and the goal is to learn which options are most rewarding within each context. After receiving complete feedback (factual and counterfactual outcomes) on trial *t*, the models update the reward expectations for the chosen and nonchosen options according to a delta rule (Rescorla & Wagner, 1972):

$$Q_{t+1}(s,i) = Q_t(s,i) + \alpha \cdot \delta_{i,t} \quad (1)$$

$$\delta_{i,t} = v_{i,t} - Q_t(s,i) \quad (2)$$

where $Q_{t+1}(s,i)$ is the updated expectation for the *i*th option in context *s* on trial *t*+1, $Q_t(s,i)$ is its previous expectation, and $\delta_{i,t}$ is the reward prediction error, or the difference between experienced ($v_{i,t}$) and expected outcomes for the *i*th option on trial *t*. The models use separate learning rates for chosen ($0 \leq \alpha_c \leq 1$)

and unchosen ($0 \leq \alpha_u \leq 1$) options to account for asymmetries in learning from factual and counterfactual outcomes. Higher values of α_c and α_u result in faster learning from recent outcomes (i.e., greater recency effects), whereas lower values result in more gradual learning.

Given the set of updated reward expectations for the *K* available options, all models use the softmax function to compute the probability of choosing the *i*th option on trial *t*+1:

$$P_{t+1}(\text{choose } i\text{th option}) = \frac{e^{\beta \cdot Q_{t+1}(s,i)}}{\sum_{k=1}^{K} e^{\beta \cdot Q_{t+1}(s,k)}} \quad (3)$$

where β is an inverse temperature parameter that modulates the sensitivity of choice probabilities to expected reward ($0 \leq \beta < \infty$). Higher values of β lead to greater exploitation of options with higher expected rewards, whereas lower values of β result in more random choices.

The key difference between models is how they encode experienced outcomes ($v_{i,t}$ in Equation 2). The most basic model is a standard Q-learning algorithm, modified to allow for counterfactual learning, which assumes that outcomes are encoded in an absolute fashion:

$$v_{i,t} = r_{i,t} \quad (4)$$

where $r_{i,t}$ is the objective reward from the *i*th option on trial *t*. This model tracks context-independent expected rewards for each option and thus has no way of accounting for context effects in RL (Bavard et al., 2018; Klein et al., 2017;

Palminteri et al., 2015). In contrast, the models discussed below assume that outcomes are encoded in a relative fashion by comparing them to either a single contextual reference point (reference point centering), the endpoints of a contextual distribution (range adaptation), or to other outcomes in the immediate or recent context (frequency encoding). These models will be compared to the Q-learning model, which serves as a common baseline, and to each other to determine which mechanism provides the best account of context effects in RL.

Reference Point Centering

Reference point theories assume that each reward is compared to a single value or reference point that summarizes the central tendency of previously experienced rewards within a particular context (Palminteri & Lebreton, 2021). This idea can be traced back to adaptation level theory (Helson, 1964), which posits that the perception of a target stimulus reflects the difference between the target and an adaptation level (AL), or a weighted average of multiple contextual stimuli. Adaptation level theory can account for why a person might judge a 4 oz fountain pen to be heavy and a 32 oz baseball bat to be light, despite the latter having a much greater objective weight (Helson, 1964): The average fountain pen weighs less than 4 oz and the average baseball bat weighs more than 32 oz. Similarly, an affectively neutral outcome like receiving zero reward can be disappointing in a context of gains and yet function as a reinforcer in the context of losses (Palminteri et al., 2015). This is because a zero outcome falls below the AL when all other outcomes are positive but above the AL when all other outcomes are negative.

In computational RL models, reference point-dependence can be accomplished by centering choice outcomes on a running estimate of the average reward in the current context. For example, the REFERENCE model (Palminteri et al., 2015; Palminteri & Lebreton, 2021) encodes outcomes as the difference between the objective reward, $r_{i,t}$, and the mean reward for the current context, $V_t(s)$, which serves as the reference point:

$$v_{i,t} = r_{i,t} - V_t(s) \tag{5}$$

where the current choice context *s* is the specific combination of options present on trial *t*. The Q values for each option in the current context are updated using these mean-centered rewards. Thus, expected rewards are adjusted upward whenever outcomes are better-than-average and downward whenever outcomes are worse-than-average for the given context. The model updates the reference point on each trial with a separate learning rate, α_V ($0 \leq \alpha_V \leq 1$):

$$V_{t+1}(s) = V_t(s) + \alpha_V \cdot \delta_{V,t} \tag{6}$$

where $\delta_{V,t}$ is a prediction error calculated as:

$$\delta_{V,t} = \frac{1}{K}\sum_{k=1}^{K} r_{k,t} - V_t(s) \tag{7}$$

Equation 7 shows that the average reward across the *K* available options on trial *t* is used to update $V_t(s)$, so that the update is independent of which option was chosen. In partial feedback contexts, counterfactual outcomes are unavailable and therefore $r_{k,t}$ is replaced with $Q_t(s,k)$ for all unchosen options in Equation 7

(see Palminteri et al., 2015 for tests of alternative specifications). The reference points $V_t(s)$ will gradually approximate the mean rewards in their corresponding contexts, with the speed of convergence determined by α_V. It is important to note that relative valuation effects in this model increase over time because they depend on learning the average reward in each context; thus, for a limited number of trials, the context dependence will only be partial. Higher (lower) values of α_V result in faster (slower) development of relative encoding. If $\alpha_V = 0$, the model reduces to the standard Q-learning model. If $\alpha_V > 0$, the option-specific Q values will gradually reflect the expected *relative* value of each option with respect to its local reference point.

Reference point-dependence has been used to explain a variety of behavioral effects in RL. Because avoided punishments carry a positive relative value in contexts where the average outcome is negative, the REFERENCE model provides a parsimonious account of avoidance learning that the standard Q-learning model fails to capture (Palminteri et al., 2015). It predicts patterns of irrational preferences that can arise when choice options whose values were learned in one context are reencountered in novel contexts (Bavard et al., 2018; Klein et al., 2017; Palminteri et al., 2015). At the same time, the REFERENCE model's partial centering mechanism that evolves over time permits the differentiation between the best outcomes in good versus bad contexts (Burke et al., 2016; see also Palminteri & Lebreton, 2021). Further, the observation of slower decision times and lower confidence in punishment compared to reward contexts (despite similar levels of accuracy) has been explained by linking

response times and confidence ratings to the mean rewards tracked by the REFERENCE model (Fontanesi et al., 2019; Lebreton et al., 2019).

Range Adaptation

Range adaptation theories assume that rewards are evaluated based on their relative position with respect to the range of rewards within a particular context (Palminteri & Lebreton, 2021; see also Volkmann, 1951). This idea has its roots in range-frequency (RF) theory (Parducci, 1965; 1995), according to which the *range value* of target stimulus S_i, denoted R_i, is computed as:

$$R_i = \frac{S_i - S_{min}}{S_{max} - S_{min}} \tag{8}$$

where S_i is the objective value of stimulus i, and S_{min} and S_{max} are the most extreme stimulus values in the judgment context. Thus, R_i can be interpreted as the proportion of the range of stimulus values that fall below the value of the target stimulus. Returning to an earlier example, a 4 oz fountain pen would have a range value close to 1 in the context of other fountain pens because 4 oz is near the top of the range; on the other hand, a 32 oz baseball bat would have a range value close to 0 in the context of other baseball bats because 32 oz is near the bottom of the range. We should then expect the pen to be judged heavy and the bat to be judged light. Similarly, zero reward would receive a range value of 0 in the context of gains, making it the least attractive outcome, but a range value of 1 in the context of losses, making it the most attractive outcome. In these

cases, the predictions of range adaptation are consistent with the predictions of reference point centering.

Bavard et. al's (2021) RANGE model incorporates a dynamic range adaptation process into an RL framework. Range-normalized outcomes are computed by subtracting the subjective minimum reward from the objective outcome and dividing by the subjective range of rewards in the current context:

$$v_{i,t} = \frac{r_{i,t} - R_{MIN,t}(s)}{R_{MAX,t}(s) - R_{MIN,t}(s)} \tag{9}$$

where $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are the current estimates of the minimum and maximum reward in context *s*. Option values are updated using the range-normalized outcomes. The estimates of the maximum and minimum rewards in context *s* are updated separately:

$$\begin{aligned} R_{MAX,t+1}(s) &= R_{MAX,t}(s) + \alpha_R \cdot \left(MaxReward_t(s) - R_{MAX,t}(s)\right) \\ R_{MIN,t+1}(s) &= R_{MIN,t}(s) + \alpha_R \cdot \left(MinReward_t(s) - R_{MIN,t}(s)\right) \end{aligned} \tag{10}$$

where $MaxReward_t(s)$ and $MinReward_t(s)$ are the objective maximum and minimum reward values observed in context *s* through the first *t* trials and α_R is a learning rate ($0 \leq \alpha_R \leq 1$). In the present application, $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are initialized to the global minimum and maximum reward values across all choice contexts. Thus, if $\alpha_R = 0$, subjective values are adapted to the *global* range of rewards and the qualitative predictions of the RANGE model are consistent with the qualitative predictions of the standard Q-learning model. If $\alpha_R > 0$, $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ gradually converge to the *local* minimum and maximum rewards in

context *s* with increasing experience. In either case, the range-normalized outcomes $v_{i,t}$ are bounded between 0 and 1 (Equation 9). As in the REFERENCE model, relative encoding strengthens over time as the subjective endpoints converge to the actual endpoints of the contextual distributions (see Palminteri & Lebreton, 2021).[1]

Range adaptation can account for similar learning performance in contexts with small and large magnitude outcomes (Bavard et al., 2018; 2021). The standard Q-learning model, in contrast, predicts stronger learning in response to larger rewards. Range adaptation facilitates learning by amplifying the gain on outcome differences in small magnitude contexts and reducing the gain in large magnitude contexts so that outcomes are experienced similarly in both. This is consistent with evidence that the firing rates of specific neurons adjust to match the range or variability of reward values within recent temporal context (e.g., Kobayashi et al., 2010; Padoa-Schioppa, 2009; Tobler et al., 2005). The RANGE model also predicts some of the irrational preferences that are observed when options are taken out of their original learning contexts: Options that were relatively better in small magnitude contexts may be preferred over options that

[1] The updating scheme in Equation 10 is slightly different than the one proposed by Bavard et al. (2021). In that study, $R_{\mathrm{MIN},t}(s)$ and $R_{\mathrm{MAX},t}(s)$ are both initialized to zero and updated only if the minimum reward on trial *t* is lower than $R_{\mathrm{MIN},t}(s)$ or the maximum reward on trial *t* is greater than $R_{\mathrm{MAX},t}(s)$. In the present study, the reward values in certain contexts were all greater than zero and thus the subjective minimum would never be updated under that scheme. Equation 10 ensures that the subjective minimum and maximum will both be updated. In any case, the asymptotic predictions of the model are very similar under both approaches.

were relatively worse in large magnitude contexts, even if the latter have a higher objective value (Bavard et al., 2018; 2021). Finally, the model's assumption that RL agents track the minimum and maximum rewards in each context is consistent with the finding of enhanced memory for the most extreme outcomes in experience-based decision tasks (Madan et al., 2014).

Frequency Encoding

There is a third computational mechanism that can explain context effects in value-based decision making, but which has received less attention in the RL literature. Theories such as decision by sampling (DbS; Stewart et al., 2006) and the "frequency principle" of RF theory (Parducci, 1965; 1995) postulate that the subjective value of a stimulus is determined by its *rank* within the contextual distribution. More specifically, DbS proposes that individuals possess two basic operations for computing subjective values: binary ordinal comparison (i.e., greater than, less than, or equal to) and frequency accumulation. When evaluating a target stimulus S_i, individuals are assumed to compare it to a sample of N stimuli that are present in the immediate context, retrieved from memory, or both. Comparisons are made by counting the number of stimuli in the sample that are less than the target. The subjective value of the target is equal to its proportional rank within the comparison sample (or, in the terminology of RF theory, its *frequency value,* F_i):

$$F_i = \frac{\text{rank}(S_i) - 1}{N - 1} \tag{11}$$

where $\mathrm{rank}(S_i)$ is the number of stimuli in the comparison sample that are less than S_i. Thus, according to the frequency principle, a target reward will have a high subjective value if it is larger than most of the other rewards in the immediate or recent context. Only the frequency of comparisons that favor the target will determine its subjective value; the magnitudes of the relative advantages and disadvantages are irrelevant.

Rank encoding is grounded in psychophysical evidence that people are better at discriminating between stimuli than estimating their absolute magnitudes (Stewart et al., 2005). Given that memories of exact stimulus values are often noisy and degraded, it may be easier to retrieve stimulus ranks (or comparative judgments) from the encoding context and use these ranks to reconstruct the stimulus values (Choplin & Hummel, 2002; Higgins & Lurie, 1983; Wedell, 1996). Some have argued that the need to store stimulus values in memory for later retrieval induces a shift to rank encoding to enhance the differences between individual stimuli (Pettibone & Wedell, 2007; Wedell, 1996). In support of this, prior research has demonstrated that when stimulus values are associated with name cues and retrieved from long-term memory using a cued recall procedure, reproduced estimates and category ratings exhibit biases that are consistent with rank encoding (Choplin & Wedell, 2014; Pettibone & Wedell, 2007; Wedell et al., 2020). Further, frequency effects are well-documented in category rating tasks that manipulate the skewing of the category distributions while holding the endpoints of the distributions constant (e.g., Niedrich et al., 2001; Parducci, 1995; Wedell, 1996; Wedell et al., 2020).

Hayes and Wedell (2021) introduced a simple RL model that implements frequency encoding on a trial-by-trial basis. The original model relied on immediate ordinal comparisons among factual and counterfactual outcomes in choice tasks with complete feedback. Here, we introduce a more general version—the FREQUENCY model—which applies to both complete and partial feedback contexts. Importantly, the FREQUENCY model allows for ordinal comparisons with previous contextual outcomes held in memory, consistent with DbS and RF theory.

Suppose that on trial *t*, an individual makes a choice between the options in context *s*. Let $\mathbf{r}(s)$ denote a vector containing all rewards observed in context *s* from trial 1 up to and including trial *t*. This vector is essentially an exemplar-based representation of the current reward context. Note that in complete feedback contexts with *K* choice options, $\mathbf{r}(s)$ will contain *K* times as many outcomes as it would in partial feedback contexts. The outcomes in $\mathbf{r}(s)$ constitute the contextual distribution for evaluating the target outcomes on trial *t*; however, due to limited working memory, only some of the outcomes in $\mathbf{r}(s)$ are recruited to form the comparison sample. The result is a recency-biased estimate of the target outcome's proportional rank within the contextual distribution. Let $r_{[j]}$ denote the *j*th outcome in $\mathbf{r}(s)$ and let $t_{[j]}$ denote the trial on which it was observed. Then the frequency value of $r_{i,t}$, the outcome from the *i*th option on the current trial, is computed by the FREQUENCY model as follows:

$$F_{i,t} = \frac{\left\{\sum_j \left[\frac{\text{sign}\left(r_{i,t} - r_{[j]}\right) + 1}{2}\right] \cdot (1-\phi)^{t-t_{[j]}}\right\} - 0.5}{\left\{\sum_j (1-\phi)^{t-t_{[j]}}\right\} - 1} \quad (12)$$

where $\left[\text{sign}\left(r_{i,t} - r_{[j]}\right) + 1\right]/2$ is an ordinal comparison function that returns 1 if $r_{i,t} > r_{[j]}$, 0.5 if $r_{i,t} = r_{[j]}$, or 0 if $r_{i,t} < r_{[j]}$. The term inside curly brackets in the numerator of Equation 12 is like the $\text{rank}(\cdot)$ function in Equation 11 that tallies the number of ordinal comparisons favoring $r_{i,t}$. The difference is that here, the comparisons are weighted such that recent outcomes will have a greater impact on the evaluation of $r_{i,t}$ than earlier outcomes. Note that the weights $(1-\phi)^{t-t_{[j]}}$ are a decreasing function of the recency parameter ϕ and the number of trials since the corresponding contextual outcome $r_{[j]}$ was observed.[2] In essence, each weight can be interpreted as the activation value for the corresponding contextual outcome on trial *t*. The longer it has been since $r_{[j]}$ was observed, the lower its activation and thus the smaller the impact of its comparison to the current trial outcomes will be. Note that in complete feedback contexts, all outcomes on the current trial have activation values of 1.0. The bracketed term in the denominator of Equation 12 serves to normalize the frequency values between 0 and 1, similar to *N* in Equation 11. Subtracting 0.5 in the numerator and 1 in the denominator factors out the comparison of $r_{i,t}$ to itself.

[2] Three different parameterizations of the model were tested: one with the constraint $\phi = (\alpha_c + \alpha_u)/2$, another with the constraint $\phi = \alpha_c$, and a third with ϕ as a free parameter. Model comparison indicated that the first parameterization was the most parsimonious. Thus, it was not necessary to include an extra parameter to account for recency effects in the computation of frequency values.

Following RF theory, the overall subjective value of the *i*th outcome on trial *t* is then a weighted combination of its range and frequency values:

$$v_{i,t} = (1 - w_F) \cdot \frac{r_{i,t} - r_{min}}{r_{max} - r_{min}} + w_F \cdot F_{i,t} \quad (13)$$

where w_F controls the relative weighting of the frequency component ($0 \leq w_F \leq 1$). Higher values of w_F result in rank-based information having a greater influence on outcome encoding. Importantly, the range values in Equation 13 are calculated using the *global* minimum and maximum rewards across all contexts, r_{min} and r_{max}. This constraint puts the range and frequency values on the same scale, aiding the interpretation of w_F, while also ensuring that any context effects predicted by the model are driven entirely by the frequency principle. In the special case that $w_F = 0$, the FREQUENCY model makes the same qualitative predictions as the standard Q-learning model because the (global) range value term in Equation 13 is a linear function of the absolute outcome $r_{i,t}$.

The FREQUENCY model is consistent with the observation that frequency information exerts a powerful influence on decisions from experience (e.g., Don & Worthy, 2021; Hayes & Wedell, 2021). For example, previous research has demonstrated a preference for options that produce the best outcomes on most occasions (Barron & Erev, 2003; Erev & Barron, 2005). Choosing these options minimizes the probability of experiencing immediate regret resulting from negative counterfactual comparisons (Ahn et al., 2012). While this type of behavior is predicted by the FREQUENCY model when $w_F > 0$, it is not

necessarily predicted by the REFERENCE or RANGE models (see Experiment 2).

Range-Frequency Encoding

The local range adaptation mechanism in the RANGE model can be combined with the frequency encoding mechanism in the FREQUENCY model by substituting $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ for r_{min} and r_{max} in Equation 13:

$$v_{i,t} = (1 - w_F) \cdot \frac{r_{i,t} - R_{MIN,t}(s)}{R_{MAX,t}(s) - R_{MIN,t}(s)} + w_F \cdot F_{i,t} \quad (14)$$

This RANGE-FREQUENCY model can account for context effects produced by either mechanism, with w_F controlling the relative weighting of the frequency component. The context-level variables $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are incrementally updated on each trial just as in the RANGE model (Equation 10). Range-frequency effects have been widely documented in psychophysical, social, and affective judgments, and many studies have reported empirical estimates of w_F close to 0.5, indicating nearly equal weighting of the range and frequency principles (e.g., Birnbaum, 1974; Choplin & Wedell, 2014; Niedrich et al., 2001; Parducci, 1968; Riskey et al., 1979; Smith et al., 1989; Tripp & Brown, 2016; Wedell & Parducci, 1988; Wedell et al., 1989).

1.2 Distinguishing among Context-Dependent Encoding Models

Previous studies have demonstrated that reference point and range adaptation RL models provide a more accurate characterization of individual choice behavior than the standard Q-learning algorithm (Bavard et al., 2018;

2021; Klein et al., 2017; Palminteri et al., 2015). However, the choice tasks that were employed in these studies were not designed to discriminate between these mechanisms and frequency or range-frequency encoding. It is therefore unclear which mechanism best describes how choice feedback is represented and integrated when option values are learned in separate contexts.

For example, consider the task used in a recent study by Bavard et al. (2018). The purpose of this study was to test whether reference point centering, range adaptation, or a combination of both mechanisms best characterized choice behavior in a two-part instrumental learning task with an initial learning phase and subsequent transfer phase. In the learning phase, eight choice options were grouped into four pairs which served as stable contexts. A summary of the task design is shown in Table 1.1.

Table 1.1 Summary of an Instrumental Learning Task from a Prior Study with Model Predictions

	Context 1		**Context 2**		**Context 3**		**Context 4**	
	Mean = 0.50 Min = 0 Max = 1.00		Mean = 0.05 Min = 0 Max = 0.10		Mean = −0.05 Min = −0.10 Max = 0		Mean = −0.50 Min = −1.00 Max = 0	
Option	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Outcome	1.00	1.00	0.10	0.10	−0.10	−0.10	−1.00	−1.00
Probability	(.75)	(.25)	(.75)	(.25)	(.25)	(.75)	(.25)	(.75)
Absolute	0.75	0.25	0.075	0.025	−0.025	−0.075	−0.25	−0.75
Reference	0.25	−0.25	0.025	−0.025	0.025	−0.025	0.25	−0.25
Range	0.75	0.25	0.75	0.25	0.75	0.25	0.75	0.25
Frequency	0.63	0.37	0.63	0.37	0.63	0.37	0.63	0.37

Note. Outcomes were presented in euros (€). Options A-H produced a nonzero outcome (reward or loss) with a certain probability (.75 or .25), otherwise zero.

The last four rows schematize the subjective values for the eight options according to absolute encoding, reference point centering, range adaptation, and frequency encoding theories. The table is adapted from Figure 1 in Bavard et al. (2018).

The four contexts instantiated a 2 × 2 factorial combination of outcome valence (reward or loss) and outcome magnitude (big/1.0 or small/0.1). One of the options in each context had a 75% probability of producing a nonzero outcome, while the other option had only a 25% probability of producing a nonzero outcome (0 otherwise). The eight options are ordered from highest (A) to lowest (H) expected value. The goal in the learning phase was to learn to choose the options that maximized rewards (Contexts 1 and 2) or minimized losses (Contexts 3 and 4). Each context was presented on 20 trials for a total of 80 trials in the learning phase. In the transfer phase, participants encountered all possible binary combinations of options, many of which had not been previously encountered, and were tasked with choosing the higher-valued option in each pair.

The last four rows of Table 1.1 schematize the subjective values of the eight options according to absolute encoding, reference point, range adaptation, and frequency encoding theories. These numbers are meant to approximate the subjective values of the options at the end of the learning phase, after acquiring sufficient experience with the task (complete feedback is assumed). Absolute encoding results in subjective values matching the context-independent expected values (EVs) of the eight options. For the theories that assume context dependence, the subjective values were calculated by substituting either mean-

centered outcomes, range-normalized outcomes, or outcome ranks in place of absolute outcomes in the calculation of each option's EV.[3] According to reference point centering, the favorable (unfavorable) options in each context acquire a positive (negative) subjective value, regardless of whether the context involves rewards or losses. However, the option values in the small magnitude contexts are closer together than the option values in the large magnitude contexts. In contrast, the range adaptation and frequency encoding theories both predict that the subjective advantage of the favorable option is the same across contexts. The result is that all three context-dependent models can account for equal learning in reward and loss contexts, but the reference point model predicts stronger learning in the large magnitude contexts while the other two models predict no effect of outcome magnitude.

Bavard et al. (2018) found that EV-maximization in the learning phase was not affected by outcome valence but was higher in the large magnitude contexts, and transfer phase preferences were strongly influenced by the favorableness of the options within their original learning contexts. Participants' choices were consistent with a hybrid RL model that incorporated reference point and range adaptation mechanisms, along with a partial weighting of absolute outcomes (see also Bavard et al., 2021, for a demonstration that the RANGE model provides a

[3] For example, the subjective value of Option A was calculated under the reference point model as .75 × (1.00 – 0.50) + .25 × (0 – 0.50) = 0.25, under the range adaptation model as .75 × [(1.00 – 0) / (1.00 – 0)] + .25 × [(0 – 0) / (1.00 – 0)] = 0.75, and under the frequency encoding model as .75 × [(rank(1.00) – 1) / (40 – 1)] + .25 × [(rank(0) – 1) / (40 – 1)] = 0.63, where the ranks were calculated with respect to all 40 outcomes in each context under the assumption of complete feedback (2 options × 20 trials per context = 40 outcomes per context).

parsimonious account of the data). However, the authors did not include a frequency encoding model in their set of candidate models. Our schematization of the model predictions shows that range adaptation and frequency encoding are confounded in this task (Table 1.1), and therefore it is not possible to distinguish between the two mechanisms (nor is it possible with the choice tasks used in other studies; e.g., Hayes & Wedell, in press; Klein et al., 2017; Palminteri et al., 2015). Additional work is needed to clarify which model provides a better description of context-dependent outcome encoding in human RL.

The aim of the current study was to measure human choice behavior in RL tasks that dissociate reference point, range adaptation, and frequency encoding models. This was accomplished in two separate experiments. Experiment 1 was primarily concerned with distinguishing between range adaptation and frequency encoding models, while Experiment 2 was designed to permit a more complete dissociation of all four candidate mechanisms at once (including reference point centering and range-frequency encoding). Thus, the current study sought to address the limitations of previous research and further elucidate the underlying computational processes that drive context dependence in human RL.

CHAPTER 2

EXPERIMENT 1

The purpose of the first experiment was to test competing theories of context-dependent encoding using a single choice task. The task was derived from a study on context effects in price perception by Niedrich and colleagues (2001; Experiment 1). It is well known that consumers judge prices by comparing them to an internal reference point (e.g., Adaval & Monroe, 2002). The aim of Niedrich et al.'s (2001) study was to determine whether the internal reference point for prices is best described by adaptation-level theory, range theory, or range-frequency (RF) theory. Participants were exposed to a sequence of airline ticket prices and rated each one on its unattractiveness. The prices were drawn from three different contextual distributions that differed on the mean price and the shape of the distribution (positive or negative skew). Participants were randomly assigned to one of the three contextual distributions in a between-subjects design. After rating the 20 prices in the distribution to which they were assigned, participants rated a set of five target prices that were common to all three distributions. Unattractiveness ratings for the target prices were higher when the targets were above the midpoint of the price range. Further, the empirical rating function was convex in the negatively skewed condition and concave in the positively skewed condition. Unattractiveness ratings were higher when there were many contextual prices below the targets. These results support

RF theory over adaptation-level and range theory and suggest that consumers judge prices by comparing them to several exemplars in recent context (Niedrich et al., 2001).

The present experiment built on Niedrich et al.'s (2001) study in several important ways. First, instead of a judgment task, we used a repeated decision-making task in which participants learned the values of several options through experience to make reward-maximizing choices. The price discounts that were used as stimuli in Niedrich et al.'s study were converted to rewards (points) in the present experiment. The rewards were produced by choice options grouped together in fixed contexts with different reward distributions. We expected to observe context dependence in the choices participants made when options were transferred out of their original learning contexts. While rating tasks can demonstrate context effects on perception, the choice task in the present study demonstrates the potential implications of these effects on economic behavior. Second, instead of assigning different participants to different contexts, we employed a within-subjects design in which participants were exposed to all three contexts. Each choice option belonged to one of the three contexts and the groupings were stable across the learning phase. Third, we tested competing theories of context dependence by fitting and comparing RL models instead of the regression-based models that were used to fit ratings in Niedrich et al. (2001).

The choice task in Experiment 1 utilized three learning contexts comprised of four options each (Table 2.1).

Table 2.1 Summary of the Instrumental Learning Task in Experiment 1

Context	Option 1	Option 2	Option 3	Option 4
NHM	NHM_{13}	**NHM_{30}**	NHM_{37}	NHM_{40}
Negative Skew, High Mean	0 (.20)	**20 (.20)**	35 (.60)	40 (1.00)
	5 (.20)	**25 (.20)**	40 (.40)	
Mean = 30	10 (.20)	**30 (.20)**		
Min = 0	15 (.20)	**35 (.20)**		
Max = 40	35 (.20)	**40 (.20)**		
Skew = −1.15				
PHM	PHM_{20}	PHM_{23}	**PHM_{30}**	PHM_{47}
Positive Skew, High Mean	20 (1.00)	20 (.40)	**20 (.20)**	25 (.20)
		25 (.60)	**25 (.20)**	45 (.20)
Mean = 30			**30 (.20)**	50 (.20)
Min = 20			**35 (.20)**	55 (.20)
Max = 60			**40 (.20)**	60 (.20)
Skew = 1.15				
PLM	PLM_{0}	PLM_{2}	PLM_{8}	**PLM_{30}**
Positive Skew, Low Mean	0 (1.00)	0 (.60)	5 (.60)	**20 (.20)**
		5 (.40)	10 (.20)	**25 (.20)**
Mean = 10			15 (.20)	**30 (.20)**
Min = 0				**35 (.20)**
Max = 40				**40 (.20)**
Skew = 1.15				

Note. Each choice option is associated with one to five different outcomes, each occurring with a specific frequency (shown in parentheses as a relative frequency). The negative skew context contains mostly larger rewards and only a few smaller rewards. The two positive skew contexts contain mostly smaller rewards and only a few larger rewards. Expected values for the choice options are shown as subscripts. The target options are shown in boldface.

On each trial of the learning phase, participants chose between two of the options in a particular context but received complete feedback from all four of the options in that context. One of the contexts (NHM) had a negatively skewed reward distribution, whereas the other two had positively skewed reward distributions with either a high (PHM) or low (PLM) mean reward. There were three “target” options that produced the same exact outcomes in each context

(NHM$_{30}$, PHM$_{30}$, and PLM$_{30}$). Participants never encountered the target options together during the learning phase since they belonged to different contexts; however, the subsequent transfer phase included repeated choices between each pair of targets (NHM$_{30}$ vs. PHM$_{30}$, NHM$_{30}$ vs. PLM$_{30}$, and PHM$_{30}$ vs. PLM$_{30}$). If outcomes are encoded on an absolute scale, the subjective representations of the three targets should be the same by the end of the learning phase and thus participants should be indifferent between them in the transfer phase. However, if outcome encoding is context dependent, then the representations of target options in separate contexts should differ by the end of the learning phase. We will demonstrate below that the REFERENCE, RANGE, FREQUENCY, and RANGE-FREQUENCY models learn distinct representations of the three target options and predict different choice patterns in the transfer phase.

2.1 Model Predictions

The REFERENCE, RANGE, FREQUENCY, and RANGE-FREQUENCY models were simulated *ex-ante* across a grid of parameter values in the task described above (REFERENCE: $\alpha_c, \alpha_u, \alpha_V \in \{.10, .15, \ldots, .50\}$, $\beta = .20$; RANGE: $\alpha_c, \alpha_u, \alpha_R \in \{.10, .15, \ldots, .50\}$, $\beta = 5$; FREQUENCY: $\alpha_c, \alpha_u \in \{.10, .15, \ldots, .50\}$, $w_F \in \{.50, .55, \ldots, .90\}$, $\beta = 5$; RANGE-FREQUENCY: $\alpha_c, \alpha_u, \alpha_R \in \{.1, .2, .3, .4, .5\}$, $w_F \in \{.3, .4, .5, .6, .7\}$, $\beta = 5$).[4] The parameter values were chosen primarily to magnify

[4] Because the exploitation-exploration tradeoff is not as relevant in complete feedback tasks, we set the inverse temperature parameter (β) to a single value in the simulations. The value of β was higher for the RANGE, FREQUENCY, and RANGE-FREQUENCY models to compensate for the fact that Q values in these models are bounded between 0 and 1.

the differences between the models; for example, the w_F parameter in the FREQUENCY model took values between .50 and .90 to emphasize the effects of frequency encoding. At the same time, we tried to approximate the parameter values reported in prior studies where possible (e.g., learning rates between .10 and .50; Bavard et al., 2018, and Palminteri et al., 2015). Results were averaged across the various parameter combinations.

Simulation results for the REFERENCE model are shown in Figure 2.1. The reference points $V_t(s)$ are initialized to 30, the midpoint of the global reward distribution, and converge across the learning phase to the mean reward in each context (Figure 2.1A). Because mean-centered rewards are linearly related to objective rewards within contexts, the model predicts that agents learn to choose the EV-maximizing options during the learning phase (Figure 2.1B). Figure 2.1C shows the evolution of the Q values for the target options across trials (for simplicity, the Q values for the other options are not shown). The average reward in the NHM and PHM contexts is 30, which is the expected payoff for NHM_{30} and PHM_{30}, and thus their Q values remain stationary at 0. On the other hand, the average reward in the PLM context is 10, resulting in the Q value for PLM_{30} increasing across trials before stabilizing at 20. The agents effectively learn that the target option's expected reward value is 20 points greater than the mean for that context. Because the Q values for NHM_{30} and PHM_{30} converge to the same value, the model predicts indifference between these options in the transfer phase (Figure 2.1D). On the other hand, the higher Q value for PLM_{30} causes it to

be strongly preferred over the other targets in the transfer phase, despite the fact that they all share the same objective value.

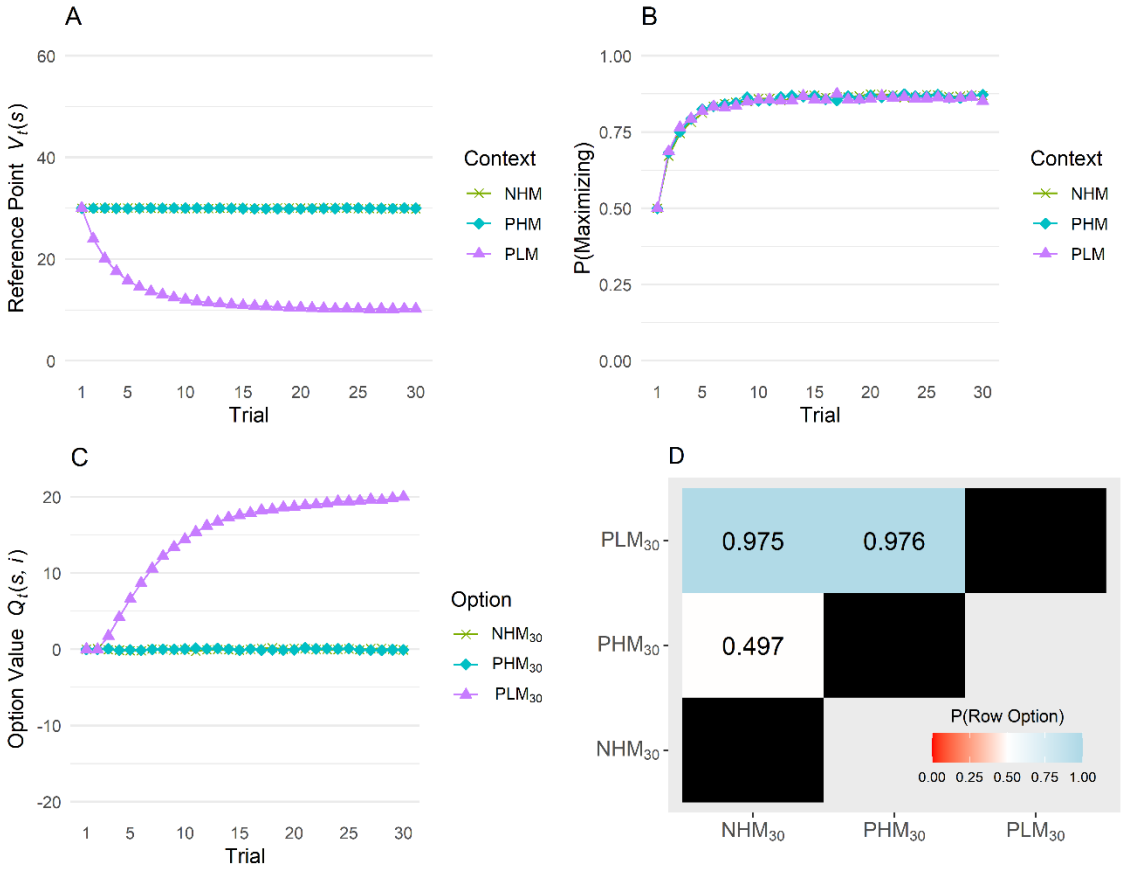


Figure 2.1 REFERENCE Model Simulations for Experiment 1. The REFERENCE model assumes that agents track a running estimate of the average reward in each context and evaluate options based on how their outcomes compare to the contextual average. (A) Learning of the reference points (i.e., average rewards) across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option Q values for the three target options across the learning phase. (D) Pairwise preferences among the three target options in the transfer phase. The numbers in each cell represent the proportion of times the row option was selected over the column option. NHM = negative skew, high mean. PHM = positive skew, high mean. PLM = positive skew, low mean.

Simulation results for the RANGE model are shown in Figure 2.2. The internal variables $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are initialized to the global minimum (0)

and maximum (60) rewards. Consequently, the subjective range for each context starts at 60 and decreases across trials before stabilizing at 40, the actual range of outcomes in all three contexts (Figure 2.2A).

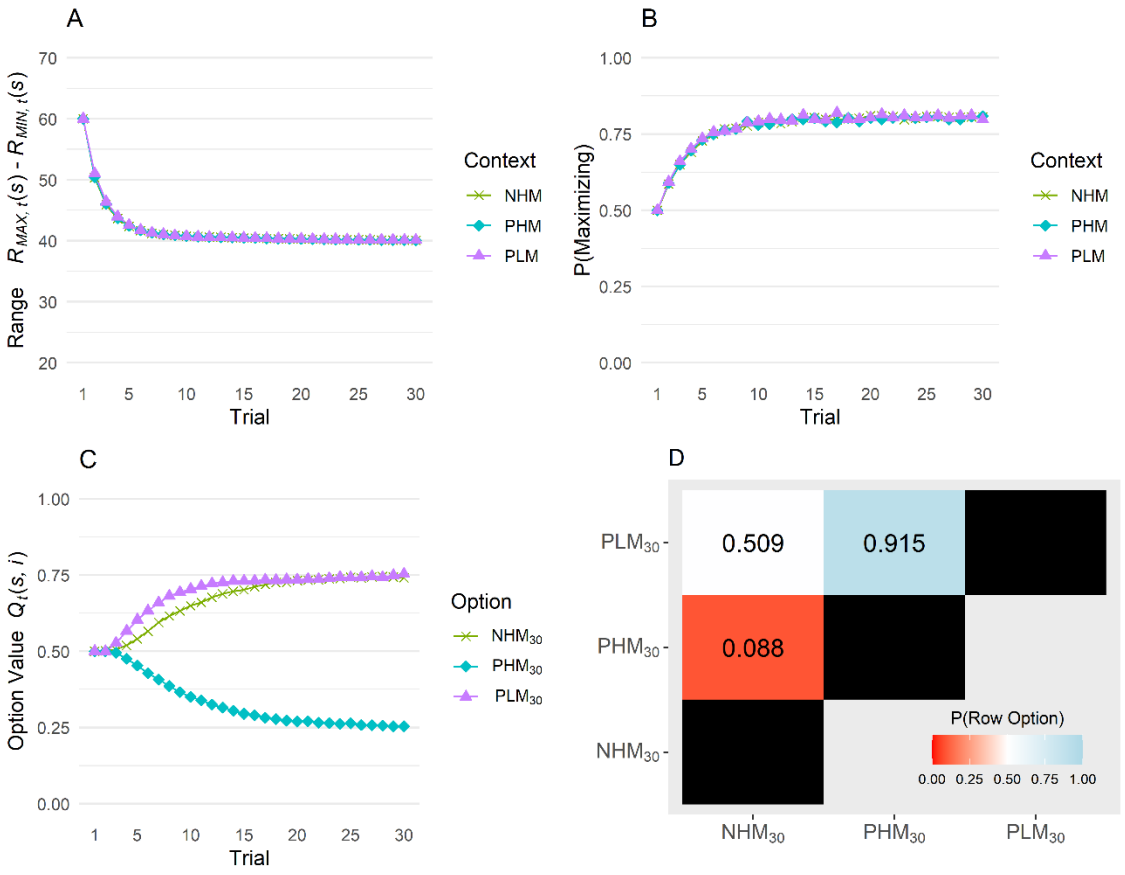


Figure 2.2 RANGE Model Simulations for Experiment 1. The RANGE model assumes that agents learn the smallest and largest rewards in each context and evaluate options based on where their outcomes fall along the contextual range. (A) Learning of the range of rewards across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option Q values for the three target options across the learning phase. (D) Pairwise preferences among the three target options in the transfer phase. The numbers in each cell represent the proportion of times the row option was selected over the column option. NHM = negative skew, high mean. PHM = positive skew, high mean. PLM = positive skew, low mean.

Because range values are linearly related to objective rewards within contexts, simulated agents learn to make EV-maximizing choices in the learning phase (Figure 2.2B). The rewards range from 0 to 40 points in the NHM and PLM contexts and from 20 to 60 points in the PHM context, but the target options always produce rewards of 20, 25, 30, 35, and 40 points. Thus, the outcomes from NHM_{30} and PLM_{30} are at or above the midpoint of the distribution (average range value = .75) whereas the outcomes from PHM_{30} are at or below the midpoint (average range value = .25). This information is reflected in the final Q values for the target options (Figure 2.2C). In the transfer phase, the result of range adaptation is an indifference between NHM_{30} and PLM_{30}, but a strong preference for both options over PHM_{30} (Figure 2.2D).

Simulation results for the FREQUENCY model are shown in Figure 2.3. There are only three panels in this figure because unlike the last two models, the FREQUENCY model does not track summary information about each context such as the mean, minimum, or maximum reward. Instead, the model maintains exemplar-based representations of each context and uses recently experienced outcomes to compute the proportional rank or frequency value of every new outcome it encounters. The weight given to frequency values during the encoding of choice feedback is determined by w_F. Like the previous models, it predicts that agents learn to make EV-maximizing choices in the learning phase (Figure 2.3A). However, the Q values exhibit a distinct progression across trials due to the reliance on frequency information (Figure 2.3B). PLM_{30} produces outcomes that are better than the outcomes of all other options in its context, and therefore its

outcomes have high frequency values. On the other hand, a substantial proportion of the outcomes in the PHM context are greater than or equal to the outcomes from PHM_{30} and an even larger proportion of the outcomes in the NHM context are greater than or equal to the outcomes from NHM_{30}. The Q values in Figure 2.3B reflect these differences in outcome ranks for the three target options.

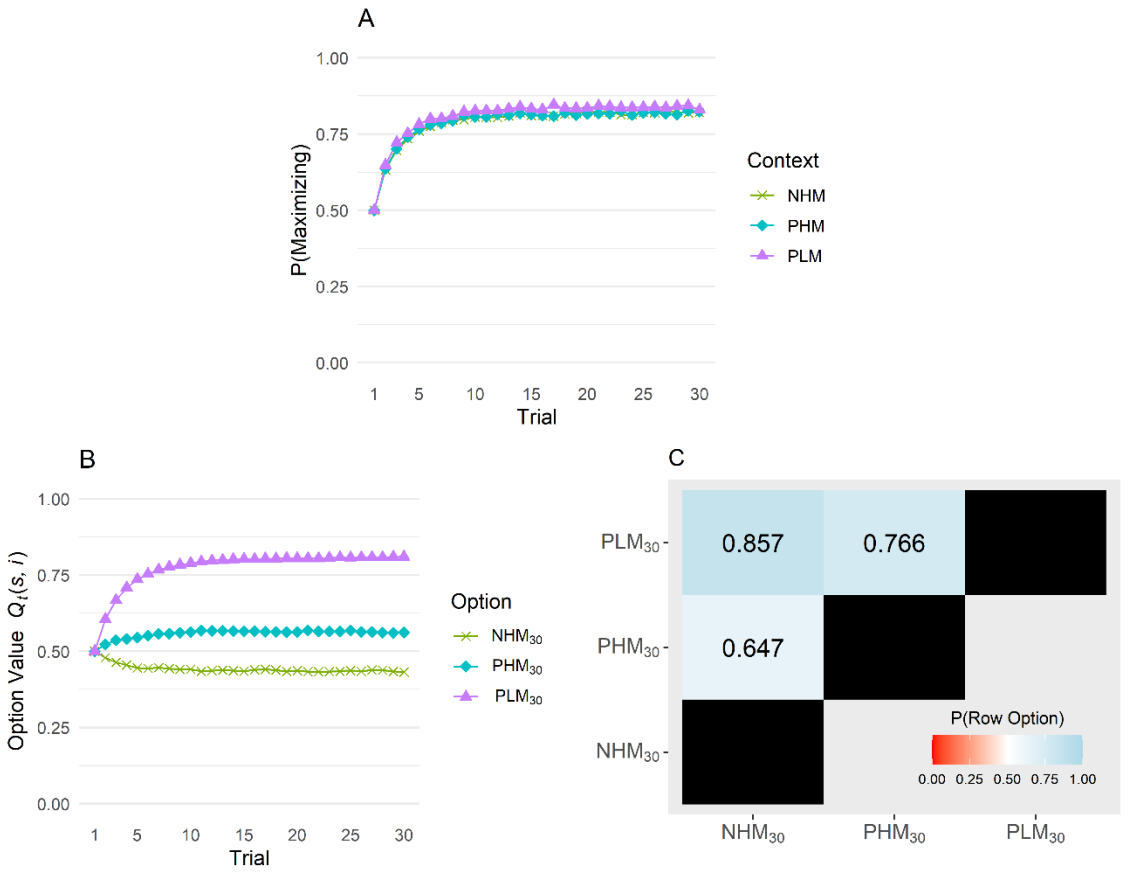


Figure 2.3 FREQUENCY Model Simulations for Experiment 1. The FREQUENCY model assumes that agents maintain exemplar representations of each context and evaluate options based on the ranks of their outcomes within the contextual distribution. (A) Predicted probability of EV-maximizing choice across the learning phase. (B) Evolution of the option Q values for the three target options across the learning phase. (C) Pairwise preferences among the three target options in the transfer phase. The numbers in each cell represent the proportion of times the row option was selected over the column option. NHM = negative

skew, high mean. PHM = positive skew, high mean. PLM = positive skew, low mean.

In the transfer phase, PLM_{30} is preferred over PHM_{30} and PHM_{30} is preferred over NHM_{30} (Figure 2.3C). Note that the FREQUENCY model's predictions conflict with the RANGE model's predictions for two of the three target pairs (NHM_{30} vs. PHM_{30} and NHM_{30} vs. PLM_{30}), making this task particularly useful for dissociating range adaptation and frequency encoding.

Finally, Figure 2.4 shows simulation results for the RANGE-FREQUENCY model. The model tracks the subjective range of outcomes in each context (Figure 2.4A) and predicts increasing maximization with increasing experience in the learning phase (Figure 2.4B). The final Q values reflect a compromise between the range and frequency components: The value for NHM_{30} is lower than the value for PLM_{30} due to the difference in outcome ranks, but higher than the value for PHM_{30} since the outcomes from NHM_{30} are above the midpoint of the contextual distribution (Figure 2.4C). Thus, the RANGE-FREQUENCY model predicts that PLM_{30} is preferred over NHM_{30} and that NHM_{30} is preferred over PHM_{30} in the transfer phase (Figure 2.4D).

The simulations above demonstrate that the RL models differ in the pattern of preferences they predict for the target choice pairs in the transfer phase. These predictions are summarized in Table 2.2. The empirical choice proportions for each of the three target pairs can be compared against chance (.50) as a test of the various context-dependent mechanisms.

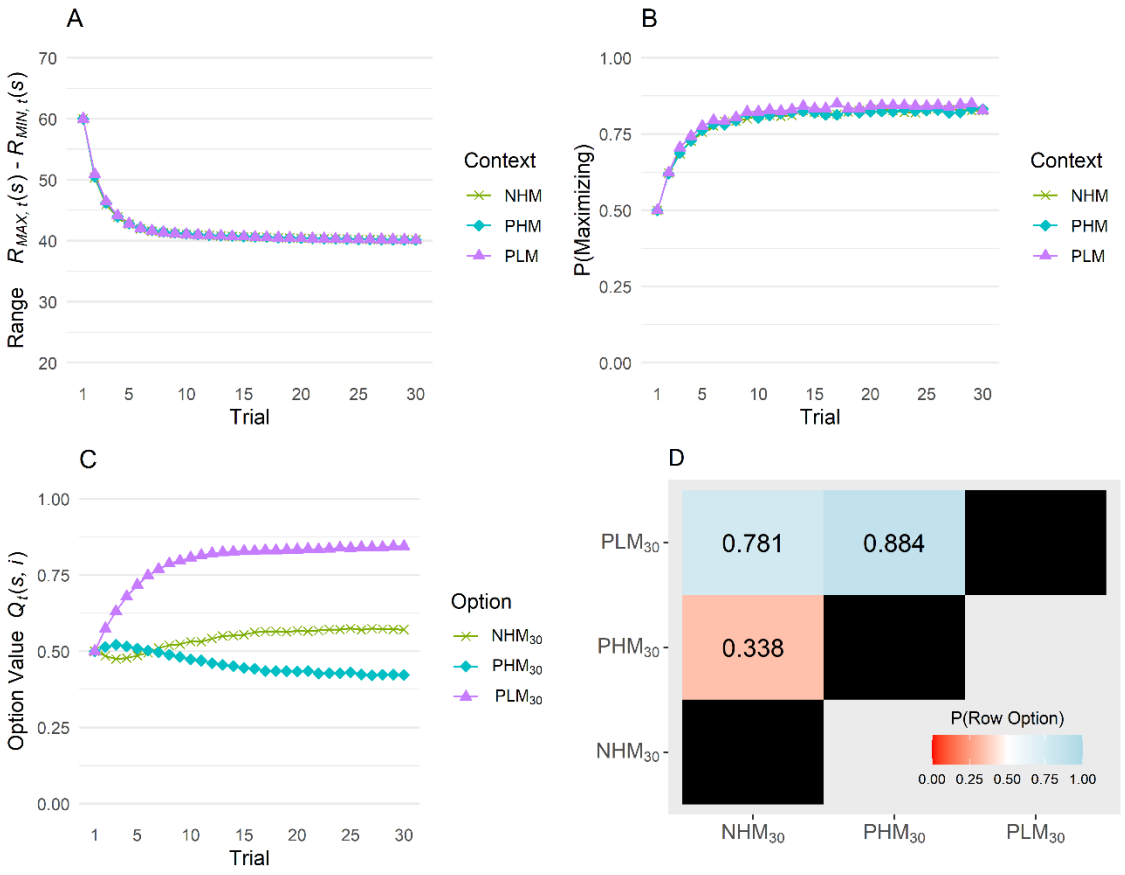


Figure 2.4 RANGE-FREQUENCY Model Simulations for Experiment 1. The RANGE-FREQUENCY model represents a compromise between range adaptation and frequency encoding models. (A) Learning of the range of rewards across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option *Q* values for the three target options across the learning phase. (D) Pairwise preferences among the three target options in the transfer phase. The numbers in each cell represent the proportion of times the row option was selected over the column option. NHM = negative skew, high mean. PHM = positive skew, high mean. PLM = positive skew, low mean.

Table 2.2 Model Predictions for the Target Choice Pairs in Experiment 1

Model	NHM_{30} vs. PHM_{30}	NHM_{30} vs. PLM_{30}	PHM_{30} vs. PLM_{30}
Q-learning	Indifference	Indifference	Indifference
REFERENCE	Indifference	PLM_{30}	PLM_{30}
RANGE	NHM_{30}	Indifference	PLM_{30}
FREQUENCY	PHM_{30}	PLM_{30}	PLM_{30}
RANGE-FREQ	NHM_{30}	PLM_{30}	PLM_{30}

Note. The cells show the preferred option predicted by each model.

2.2 Method

Our recruitment methods, experimental design, procedures, and data analysis plans were preregistered on the Open Science Framework (https://osf.io/xpn5g).

Participants

We used the crowdsourcing platform Prolific to recruit 60 participants (21 men, 36 women, 3 non-binary; ages 18 – 65, M = 35.55, SD = 12.15) for an online experiment that was administered via Qualtrics. Our inclusion criteria were age (18 – 65), nationality (US), and Prolific approval rating (at least 75%). The sample size was based on the planned analysis of the target pairs in the transfer phase (testing the null hypothesis that the choice proportions are equal to .50; see "Data Analysis and Modeling" section). To detect a medium-sized effect with .90 power, 44 participants would be required (one-sample t-test, two-tailed, d = .50, α = .05). It took participants just over 27 minutes on average to complete the experiment. Participants were informed that the points they earned in the task would be converted proportionately to real money and added to their participation

payment, although they were not informed of the conversion rate (100 points = $0.04; mean bonus = $1.90). Participants provided informed consent and all aspects of this study were approved by the Institutional Review Board at the University of South Carolina.

Design

Learning Phase. The instrumental learning task in Experiment 1 consisted of a learning phase and transfer phase. The learning phase included 12 choice options organized into three groups of four, with the groups functioning as stable choice contexts (see Table 2.1). Three of the 12 options (the "targets") produced the same outcomes and had the same EV but belonged to separate contexts (NHM_{30}, PHM_{30}, and PLM_{30}). On each trial of the learning phase, the cues for the four options in a particular context appeared on screen but only two of the options were available to choose. For example, if the NHM context was active on a particular trial, the cues for all four of its options would have appeared but the participant may have been forced to choose between NHM_{13} and NHM_{30}. This was done to encourage participants to learn the values of all four options in each context and not just the best option. With four options per context, there were $\binom{4}{2}$ = 6 possible choice pairs, and each pair was repeated five times. In total, the learning phase contained 92 trials (3 contexts × 6 choice pairs × 5 repetitions, plus 2 attention check trials). Complete outcome feedback was provided from all four options on every trial to encourage context-dependent encoding (Bavard et al., 2018; Palminteri et al., 2015). The order of context presentations was randomly interleaved for each participant. The choice cues for

all four options in a context had the same color but different shapes, and the assignment of cues to the 12 options was randomized for each participant. Using the same color for all cues within a context should make the task structure more salient and consequently enhance context-dependent encoding (Bavard et al., 2018).

Transfer Phase. The transfer phase consisted of choices between six specific cross-context pairs of options without feedback. The most diagnostic choice pairs for distinguishing between models are those formed by the three target options: NHM_{30} vs. PHM_{30}, NHM_{30} vs. PLM_{30}, and PHM_{30} vs. PLM_{30}. We will refer to these as the "target pairs." In addition, three other choice pairs were included for which at least two of the models make divergent predictions: NHM_{13} vs. PLM_{2}, NHM_{30} vs. PLM_{8}, and NHM_{37} vs. PLM_{30}. We will refer to these as the "opposite skew pairs." The opposite skew pairs are especially informative for distinguishing between the RANGE and FREQUENCY models: Range values tend to favor intermediate options from negatively skewed contexts, while frequency values tend to favor intermediate options from positively skewed contexts. The three target pairs and three opposite skew pairs were each repeated 15 times. In total, the transfer phase contained 92 trials (6 choice pairs × 15 repetitions, plus 2 attention check trials). Trial order was shuffled for each participant and options appeared an equal number of times on the left and right side of the screen.

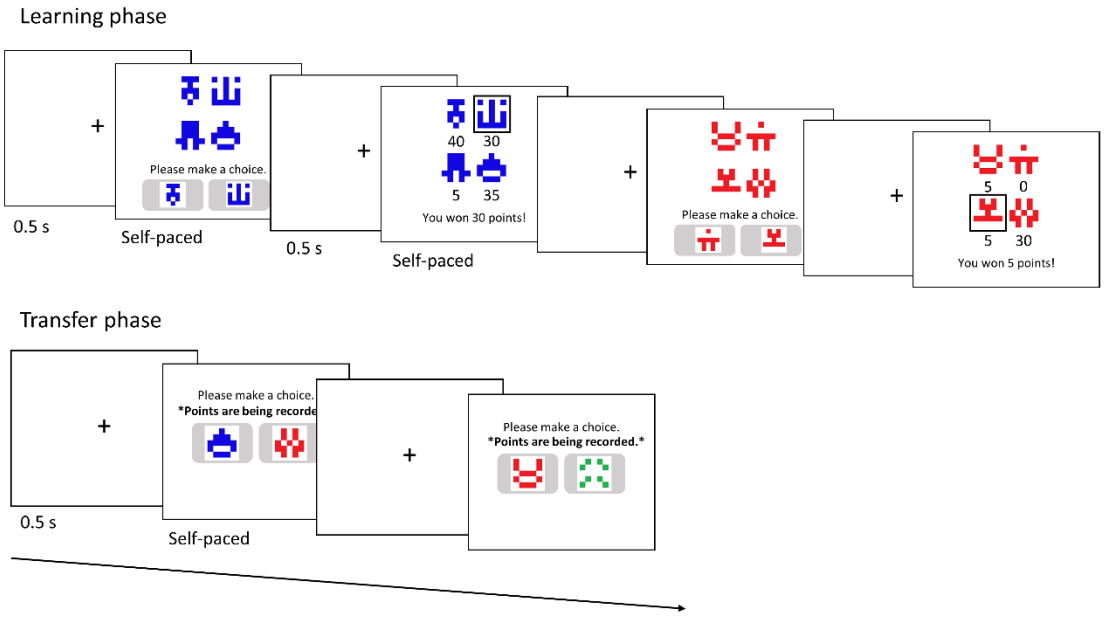


Figure 2.5 Trial Timeline for Experiment 1. Choice options were represented by cues (random identicons) that were the same color (red, green, or blue) for all options within a context. In the learning phase, each trial began with a choice prompt in which all four options in a particular context were visible but only two were available to choose. The screen locations of the four options were randomized trial-to-trial. Following the participant's choice and a 0.5 s fixation, complete feedback was presented from all four options (including those that were not selectable). The chosen option was indicated with a black border. Context presentations were randomly interleaved across trials. In the transfer phase, participants chose between specific pairs of options from different contexts without receiving feedback.

Procedure

Learning Phase. The instructions for the learning phase informed participants that they would be making repeated choices with the goal of gaining as many points as possible. Each choice that they made would result in points, but some options are more rewarding than others. Participants were told that on each trial they would see four options, but only two would be available to choose. After making a choice, they would get to see the outcomes produced by all four

options. The points from the chosen option would be added to their total (the running total was not visible).[5] They were told that the experiment contained two parts and that both parts must be completed in one sitting. Participants were not given any specific details about the transfer phase, nor were they explicitly told that there were three contexts—instead, they learned the contextual structure of the task through direct experience.

Each trial began with a 0.5 s fixation followed by the presentation of four option cues in a 2-by-2 arrangement (Figure 2.5). The four cues had the same color but different shapes, and their locations were randomized on each presentation. Two response buttons appeared at the bottom of the display that contained the cues of the two options that were available to choose on that trial. Participants indicated their choice by clicking one of the two buttons with their cursor. Following another 0.5 s fixation, participants received complete feedback from all four options presented in the same 2-by-2 arrangement, with the chosen cue indicated by a black border. The number of points produced by each option appeared just below the cue. Choices and feedback viewing were self-paced.

Transfer Phase. The transfer phase instructions informed participants that they would be making repeated choices between two options at a time without feedback, and that some of the pairs may not be ones that they had previously seen. They were told that the program would record the number of

[5] Participants were shown an example trial in which the most extreme outcomes were 0 and 60 points (i.e., the global minimum and maximum rewards). This was done to justify our choice of initial values for the RL models.

points they won from the chosen options and that they should strive to finish with as many points as possible.

Each trial began with a 0.5 s fixation followed by the presentation of two option cues arranged horizontally on screen with the message, “Please make a choice.” A reminder message stating that points are being recorded was visible at the top of the screen (Figure 2.5). Choices were self-paced.

Attention Checks. There were four attention checks that occurred at random points during the task, two in the learning phase and two in the transfer phase. The attention checks presented two symbols with different colors that had not been previously encountered, along with the instructions, "Choose the option on the [RIGHT / LEFT]." If the participant's choice did not match the instructed choice, the trial was considered a failed attention check. Participants were excluded from the analyses if they failed more than one attention check.

Post-Task Questions. After completing the transfer phase, participants answered a set of questions that further probed what they learned about the choice options. First, they were shown the 12 option cues in a randomized order and asked to rank them from highest to lowest value. Participants could rearrange the option cues using a drag and drop interface until they were satisfied that the cues were in the correct order. Second, participants were shown the four cues that belonged to each of the original learning contexts and asked to estimate the (1) average, (2) lowest, and (3) highest outcomes produced by the four options. The presentation order of the three contexts was

randomized, and participants responded using a slider ranging from 0 to 60 points. These questions were included so that we could gauge the extent to which participants maintained accurate representations of the mean, minimum, and maximum reward values in each context. The estimation questions came after the ranking question to avoid context-level estimates biasing the ranking of the individual options.

Data Analysis and Modeling

For the learning phase data, a repeated-measures analysis of variance (ANOVA) was used to analyze the proportion of EV-maximizing choices as a function of Context and Block (five blocks of six trials per context). In this task, the EV-maximizing choice on each trial was the option with a higher average payoff out of the two that were available to choose. Based on the *ex-ante* model simulations, we did not expect any effects of Context in the learning phase; however, we did expect a significant effect of Block that reflects increasing maximization with increasing experience. For the transfer phase data, we compared the choice proportions for the three target pairs (NHM_{30} vs. PHM_{30}, NHM_{30} vs. PLM_{30}, and PHM_{30} vs. PLM_{30}) against chance (.50) using one-sample t-tests. Our *ex-ante* model simulations showed that the models predict distinct patterns of preference across these choice pairs (see Table 2.2).

To elucidate the underlying computational mechanisms that may be guiding choice behavior in this task, we fit the Q-learning model and the four context-dependent RL models to each participant's choice data using maximum

likelihood methods. The models were compared based on their out-of-sample predictive accuracy in the transfer phase (for a similar approach, see Bavard et al., 2021). This process involves fitting the model to a subset of the choices for a given individual (the training data) and using the best-fitting parameters to compute the log-likelihood of the remaining choices for that individual (the test data). The best model is the one that assigns the highest log-likelihood to the test data. For each participant, the out-of-sample prediction was performed in six iterations. Each iteration involved training the model on choices in the learning phase and five out of the six transfer pairs. Then, the out-of-sample log-likelihood was computed for the remaining transfer pair. This process was repeated for each of the six pairs and the log-likelihoods were summed across iterations. In addition to the relative model comparisons, we attempted to falsify specific models by demonstrating that they were unable to generate the observed choice patterns even after conditioning on the optimized parameters (Palminteri, Wyart, et al., 2017). Code for reproducing the analyses is available at https://osf.io/br3fq/.

2.3 Results

Learning Phase

Figure 2.6A shows the proportion of EV-maximizing choices across the 30 learning trials for each context. On any given trial, the EV-maximizing choice was the option with a higher expected payoff out of the two that were available to select. Participants chose the maximizing options with increasing frequency as

they gained experience in the task, and the rate of learning did not appear to differ between contexts.

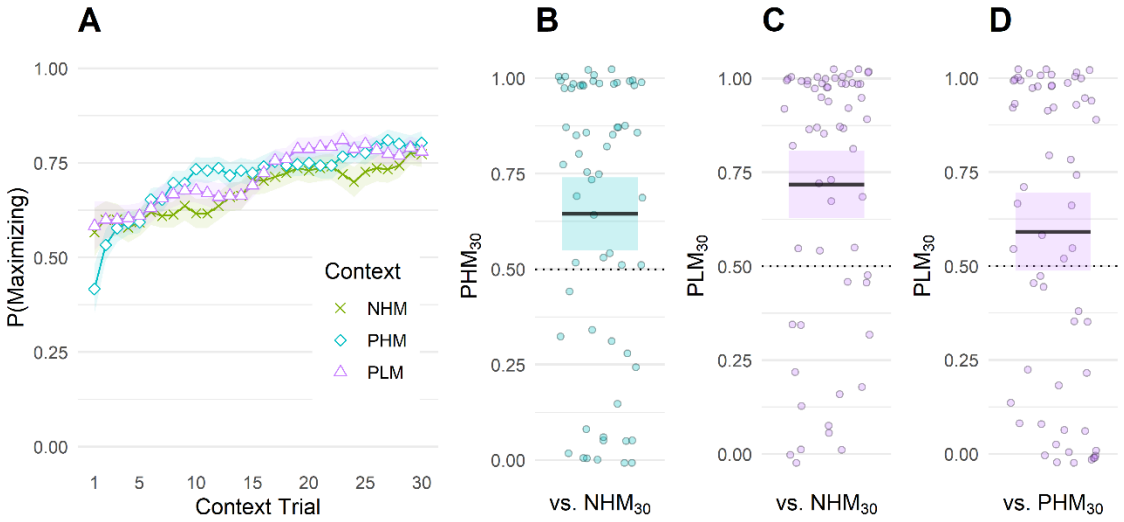


Figure 2.6 Learning and Transfer Phase Results for Experiment 1. (A) Mean proportion of EV-maximizing choices across the 30 learning trials for each context. Choices were smoothed at the individual level using a 5-trial rolling average prior to averaging across individuals. Error bands represent +/− 1 standard error. Panels B-D show the proportion of times PHM_{30} was chosen over NHM_{30} (B), PLM_{30} was chosen over NHM_{30} (C), and PLM_{30} was chosen over PHM_{30} (D) in the transfer phase. Each of these choice pairs was presented 15 times total in the transfer phase. The points show individual choice proportions, and the solid black lines show the group means. The shaded boxes show 95% confidence intervals.

To analyze the data, the 30 trials for each context were first divided into five equal-sized blocks of six trials and the proportion of maximizing choices within blocks was computed for each participant. The choice proportions were submitted to a 3 (Context) × 5 (Block) repeated-measures ANOVA. The main effect of Block was significant, $F(4, 236) = 24.02$, $p < .001$, $\varepsilon = .84$, $\eta_p^2 = .29$, confirming that the rate of maximizing choices increased over time. Follow-up polynomial contrasts revealed a significant positive linear trend across blocks

(contrast coefficient = 0.43), $t(59) = 7.53$, $p < .001$, as well as a negative quadratic trend (contrast coefficient = −0.10), $t(59) = 2.24$, $p = .029$. The main effect of Context, $F(2, 118) = 2.34$, $p = .10$, $\varepsilon = .99$, $\eta_p^2 = .04$, and the interaction, $F(8, 472) = 1.04$, $p = .40$, $\varepsilon = .74$, $\eta_p^2 = .02$, were both nonsignificant. Collapsing across blocks, the mean proportion of maximizing choices was significantly above chance in all three contexts (NHM: $M = .69$, 95% CI = [.64, .74]; PHM: $M = .73$, 95% CI = [.69, .77]; PLM: $M = .72$, 95% CI = [.68, .77]). Overall, the learning phase results were consistent with both absolute and context-dependent RL models.

Transfer Phase

The transfer phase was designed to distinguish between competing models of outcome encoding in RL. The most diagnostic choices were those that involved two of the target options (NHM_{30} vs. PHM_{30}, NHM_{30} vs. PLM_{30}, PHM_{30} vs. PLM_{30}). Absolute encoding models predict indifference for these choice pairs because all three target options produced the exact same absolute outcomes during the learning phase. In contrast, context-dependent models predict distinct preference relations among the target options, depending on how their outcomes compared to other outcomes in their respective encoding contexts (see Table 2.2).

The results for the three target choice pairs are shown in Figure 2.6B−D. Each point represents the proportion of times a single individual chose the option on the vertical axis over the option on the horizontal axis out of 15 opportunities.

The mean choice proportion (solid line) and 95% CI (shaded box) are superimposed in each panel. Although the individual preferences were somewhat noisy, the results were clearly more consistent with context-dependent encoding than with absolute encoding. First, PHM_{30} was significantly preferred over NHM_{30}, as indicated by a one-sample *t*-test on the proportion of PHM_{30} selections compared to chance (M = .64), $t(59) = 3.02$, p = .004, d = 0.39, 95% CI = [.55, .74] (Figure 2.6B). This result was also observed at the individual level: 70% of participants (n = 42) chose PHM_{30} more often than they chose NHM_{30}. Importantly, the FREQUENCY model was the only model that predicted a preference for PHM_{30} over NHM_{30} in our *ex-ante* simulations. Second, PLM_{30} was also significantly preferred over NHM_{30} on average (M = .72), $t(59) = 4.84$, $p < .001$, d = 0.63, 95% CI = [.63, .81] (Figure 2.6C), and 73% of participants (n = 44) chose PLM_{30} more times than they chose NHM_{30}. This result is consistent with our *ex-ante* simulations of the REFERENCE, FREQUENCY, and RANGE-FREQUENCY models, but inconsistent with the RANGE model. Third, PLM_{30} was preferred over PHM_{30} on average (M = .59), but the effect was only marginally significant, $t(59) = 1.76$, p = .08, d = 0.23, 95% CI = [.49, .69] (Figure 2.6D). A preference for PLM_{30} over PHM_{30} was observed for 60% of participants (n = 36). Note that all four of the context-dependent RL models predicted a preference for PLM_{30} in the *ex-ante* simulations. Taken together, the pattern of preferences that we observed for the target choice pairs most closely aligned with the predictions of the FREQUENCY model.

Model Comparison

We fit a Q-learning model, which assumes absolute outcome encoding, as well as the four context-dependent encoding models to each participant's data using maximum likelihood methods. The models were compared on two metrics: out-of-sample prediction in the transfer phase and BIC (see Method). The results of the relative model comparison are shown in Table 2.3.

Table 2.3 Model Comparison Results in Experiment 1

Model	Parameters	Out-of-sample log-likelihood (Transfer phase only)	BIC (Both phases)
Q-learning	3	−70.61*	214.31***
REFERENCE	4	−71.89*	197.72
RANGE	4	−70.29**	215.62***
FREQUENCY	4	**−59.26**	**195.46**
RANGE-FREQ.	5	−61.32	196.38

Note. Mean out-of-sample log-likelihood and Bayesian information criterion (BIC) values. The best model according to each metric is shown in bold. Significance tests reflect comparisons of each model to the best model using paired t-tests (df = 59). BIC = $-2 \times LL + k \times \ln(n)$, where LL is the maximized log-likelihood, k is the number of model parameters, and n is the number of observations.
*$p < .05$, **$p < .01$., ***$p < .001$

The FREQUENCY model had the highest mean out-of-sample log-likelihood and lowest mean BIC across participants, indicating that it was the best model overall. Paired t-tests revealed that the FREQUENCY model was significantly better than the Q-learning and RANGE models according to both metrics, and significantly better than the REFERENCE model according to out-of-sample prediction in the transfer phase. However, the mean estimate of the frequency weight (w_F) in our sample was .40, which was lower than the range of values used in the *ex-ante* simulations (see Table 2.4 for a full summary of the parameter estimates). The RANGE-FREQUENCY model, which combines range

adaptation and frequency encoding mechanisms, did not lead to an improvement over the simpler FREQUENCY model and was likely disadvantaged by having an extra parameter.

Table 2.4 Mean Parameter Estimates in Experiment 1

Model	β	α_c	α_u	α_V	α_R	w_F
Q-learning	2.58 (6.45)	.40 (.42)	.37 (.39)	--	--	--
REFERENCE	3.98 (7.45)	.31 (.37)	.29 (.37)	.22 (.36)	--	--
RANGE	9.04 (7.86)	.41 (.43)	.35 (.38)	--	.17 (.32)	--
FREQUENCY	11.48 (8.09)	.31 (.36)	.27 (.37)	--	--	.40 (.37)
RANGE-FREQUENCY	12.10 (8.01)	.28 (.36)	.25 (.35)	--	.24 (.35)	.38 (.36)

Note. Standard deviations shown in parentheses. β = inverse temperature, α_c = chosen learning rate, α_u = unchosen learning rate, α_V = reference point learning rate, α_R = range learning rate, w_F = frequency value weighting.

The models were also assessed on their ability to reproduce the observed choice patterns using the parameter values obtained from model fitting. Each model was simulated 100 times in the task using each participant's optimized parameters and the predicted choice probabilities were averaged over iterations. Note that the models were not provided with participants' actual choice histories for the simulations, making this a test of generative performance (Palminteri, Wyart, et al., 2017; Steingroever et al., 2014). Because all the models predict similar behavior in the learning phase, we focused on their ability to reproduce the choice patterns in the transfer phase. The results are shown in Figure 2.7.

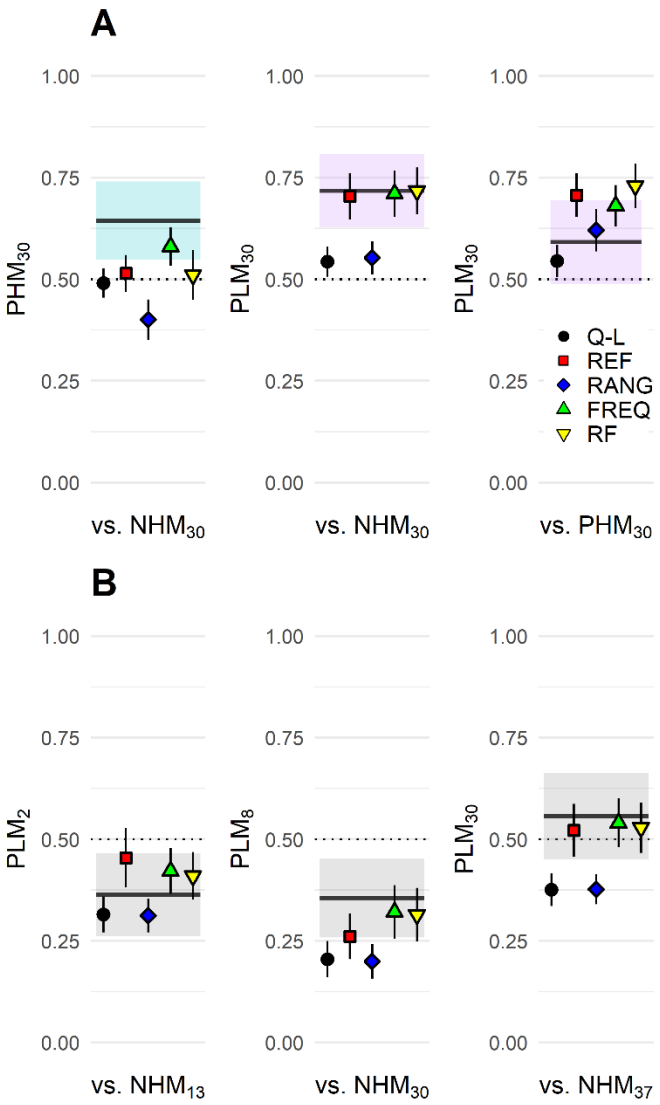


Figure 2.7 Transfer Phase Pairwise Choice Preferences in Experiment 1: Empirical Data vs. Model Simulations. Each panel shows the mean proportion of times the option on the vertical axis was chosen over the option on the horizontal axis, averaging across participants. Target pairs are shown in (A), opposite skew pairs in (B). The solid black lines and shaded boxes show the means and 95% confidence intervals for the observed data. The points and error bars show the means and 95% confidence intervals for the RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. Q-L = Q-learning. REF = REFERENCE. RANG = RANGE. FREQ = FREQUENCY. RF = RANGE-FREQUENCY.

The FREQUENCY model reproduced the pairwise preferences among the target options more accurately than the other models (Figure 2.7A). Critically, it was the only model that captured the significant preference for PHM_{30} over NHM_{30}. The REFERENCE model could not account for this effect since both options had expected payoffs that were equal to the mean rewards in their respective encoding contexts. The RANGE model predicted a significant preference in the *opposite* direction because the outcomes from NHM_{30} had higher range values than the outcomes from PHM_{30}. The only advantage for PHM_{30} over NHM_{30} is that its outcomes had higher frequency values. In addition, both the Q-learning model and the RANGE model greatly underestimated the observed preference for PLM_{30} over NHM_{30}, whereas the other models reproduced this effect accurately. The FREQUENCY model was also capable of reproducing the transfer preferences for the opposite skew pairs (Figure 2.7B). In contrast, the Q-learning and RANGE models considerably overestimated selections of NHM_{30} over PLM_{8} and predicted a significant preference for NHM_{37} over PLM_{30}, even though the average participant did not exhibit a significant preference for either option.

Overall, the relative model comparison and *ex-post* simulations falsify the basic Q-learning model and favor the FREQUENCY model as the best explanation of the context-dependent choice behavior that we observed. However, an alternative interpretation of our results is that individuals develop habits of choosing certain options more frequently than other options during the learning phase, and that differences in habit strengths might explain preferences

in the transfer phase. If this is the case, we would expect a positive association between the number of times an option was chosen during learning and during transfer. We calculated the number of times that each of the target options was chosen during the learning phase as an indicator of habit strength. Then, for each pair of target options, we correlated the *difference* in habit strengths with the proportion of times one option was chosen over the other in the transfer phase. It should be noted that finding a positive correlation would not rule out context-dependent value encoding as a contributing mechanism, but it would suggest a possible role of stimulus-response associations (i.e., habits) alongside stimulus-outcome associations (i.e., subjective values) in guiding choice behavior. There was a positive association between the relative habit strengths for PLM_{30} versus NHM_{30} and the proportion of times PLM_{30} was chosen over NHM_{30} in the transfer phase, $r(58) = 0.43$, $p < .001$. In other words, the more that PLM_{30} was chosen during the learning phase compared to NHM_{30}, the more it was chosen over NHM_{30} during transfer. However, the correlations for the other two pairs of target options were nonsignificant [NHM_{30} vs. PHM_{30}: $r(58) = -0.04$, $p = .77$; PHM_{30} vs. PLM_{30}: $r(58) = 0.15$, $p = .25$]. These results suggest a minimal role of habitual processes in this paradigm, consistent with prior work (Bavard et al., 2021). Critically, habit strengths cannot account for the significant preference for PHM_{30} over NHM_{30} that was exclusively predicted by the FREQUENCY model.

Post-Task Questions

After completing both phases of the choice task, participants were shown the 12 option cues in a randomized order and asked to rank them from highest (1) to lowest (12) value using a drag and drop interface. The mean reported ranks for each option are shown in Figure 2.8A. Participants' rankings were somewhat sensitive to the absolute values of options, but contextual biases were still present. For example, the average ranks for the target options—which should be equal under absolute encoding—followed the pattern predicted by the FREQUENCY model ($NHM_{30} < PHM_{30} < PLM_{30}$). Thus, the rank judgments were generally consistent with the pattern of choice preferences observed in the transfer phase.

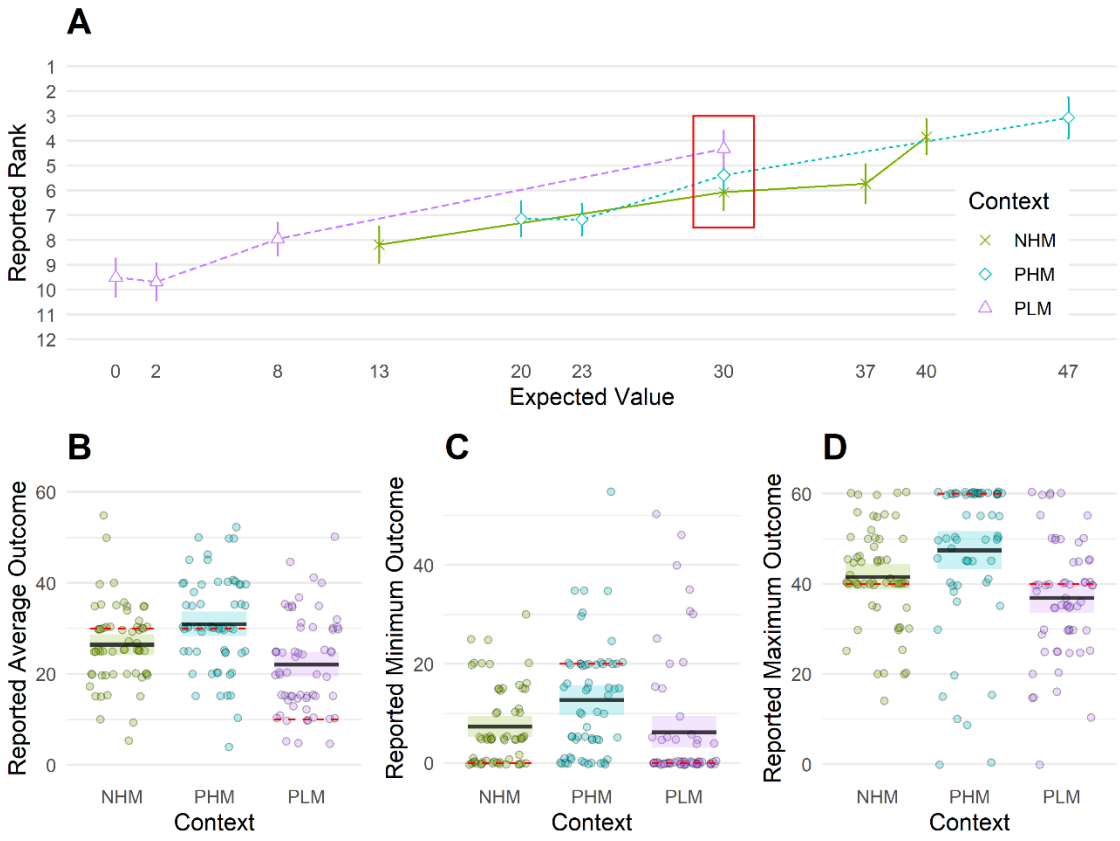


Figure 2.8 Responses to the Post-Task Questions in Experiment 1. (A) Mean reported ranks for each choice option. Error bars show 95% confidence intervals. The box contains the mean ranks for the three target options. Panels B-D show the reported average (B), minimum (C), and maximum outcomes (D) for each context. Individual estimates are shown as points and the group means are shown as solid black lines. Shaded boxes represent 95% confidence intervals. The true average, minimum, and maximum outcomes are shown as dashed red lines.

Participants were also asked to estimate the (1) average, (2) lowest, and (3) highest outcomes produced by the four options in each context considered as a group. They responded using a slider that ranged from 0 to 60 points. The individual-level estimates and group averages are shown in Figure 2.8B−D. Clearly, the individual estimates were quite noisy, and the aggregate estimates often deviated significantly from the actual values of the outcome statistics. Nevertheless, participants appeared to be somewhat sensitive to the relative differences between contexts. For example, most participants seemed to recognize that the lowest outcome in the PHM context (20) was higher than the lowest outcomes in the NHM and PLM contexts (0). Some of the extreme outcomes appeared to be recalled accurately by a considerable number of participants, especially when those outcomes were produced by one of the four options on every trial (e.g., PLM_0 always produced 0 points).

Three separate ANOVAs were run on the estimation data—one for each outcome statistic (average, minimum, maximum)—with Context as a repeated measures factor. These analyses were not preregistered and should be considered exploratory. The effect of Context was significant in all three cases [average: $F(2, 118) = 18.97$, $p < .001$, $\varepsilon = .98$, $\eta_p^2 = .24$; minimum: $F(2, 118) =$

7.14, p = .001, ε = .92, η_p^2 = .11; maximum: $F(2, 118) = 10.07$, $p < .001$, ε = .94, η_p^2 = .15]. Participants reported the highest average outcome for the PHM context, followed by the NHM and PLM contexts (Figure 2.8B). Pairwise comparisons (Tukey-corrected) indicated that the differences between the three context pairs were significant (all *p*s < .013), which does not match the actual pattern of average outcomes across contexts (i.e., the NHM and PHM contexts had the same average outcome). Participants reported the highest minimum outcome for the PHM context, followed by the NHM and PLM contexts (Figure 2.8C). The differences between PHM and NHM (p = .003) and between PHM and PLM (p = .006) were both significant, but the difference between NHM and PLM was not (p = .82), matching the actual pattern of minimum outcomes across contexts. Finally, participants reported the highest maximum outcome for the PHM context, followed by the NHM and PLM contexts (Figure 2.8D). The differences between PHM and NHM (p = .032) and between PHM and PLM (p < .001) were both significant, but the difference between NHM and PLM was not (p = .08), matching the actual pattern of maximum outcomes across contexts.

2.4 Discussion

Experiment 1 used a choice task which was adapted from a previous study on price perceptions to investigate the nature of context-dependent outcome encoding in RL. In that study (Niedrich et al., 2001), the authors demonstrated that RF theory provided a better explanation of context dependence in price judgments than adaptation-level and range theories. The results of the present study were generally consistent with Niedrich et al. and

extend their conclusions to the domain of experience-based choice. We showed that an RL model inspired by theories such as DbS (Stewart et al., 2006) and RF theory (Parducci, 1965, 1995) was the best description of participants' choice behavior out of the models we considered (the FREQUENCY model).

In the transfer phase, there was a significant preference for the target options from positively skewed reward contexts (PHM_{30} and PLM_{30}) over the target option from a negatively skewed reward context (NHM_{30}) even though all three options produced the same absolute outcomes during the learning phase. This result clearly falsifies RL models which assume absolute outcome encoding, such as the standard Q-learning model. Moreover, we demonstrated that the transfer preferences were inconsistent with RL models that implement context dependence using dynamic adaptation-level (REFERENCE model) or range adaptation mechanisms (RANGE model). The results were best explained by a frequency encoding mechanism which assumes that outcomes are compared to other outcomes in an ordinal fashion to determine their rank within the local contextual distribution. The FREQUENCY model outperformed the others both in terms of relative model comparison criteria (out-of-sample prediction and BIC) as well as generative performance (*ex-post* simulations). The RANGE-FREQUENCY model, which combines the frequency encoding and dynamic range adaptation mechanisms, was a close second; however, the additional range adaptation parameter was not necessary to account for choice behavior in this experiment. Further, we showed that the observed choice behavior could not be adequately explained by the hypothesis that decision makers are guided

solely by stimulus-response associations (i.e., habits), as opposed to subjective values.

Lastly, participants' rankings of the options at the end of the task were mostly consistent with the aggregate choice preferences in the transfer phase. This shows that context dependence in RL is observable using other elicitation methods beyond choice (Soukupová et al., 2021). Participants were not very accurate when asked to recall the average, minimum, and maximum outcomes in each learning context at the end of the task. However, it is certainly possible that they may have forgotten this information gradually over the course of the transfer phase and that recall accuracy would have been higher if the probes had occurred immediately after the learning phase. Nevertheless, participants' estimates of the highest and lowest outcomes in each context matched the structure of the task, even though the estimates significantly deviated from the actual values. This may be due to enhanced memory for outcomes at or near the edge of the contextual distribution (Madan et al., 2014).

CHAPTER 3

EXPERIMENT 2

As in the first experiment, the purpose of Experiment 2 was to test competing theories of context-dependent RL using a single choice task for which the theories make distinct predictions. However, the choice task in the second experiment was designed to provide a more powerful test of the candidate models. During the learning phase, eight choice options were presented in fixed pairs to encourage context-dependent encoding: The same two options were always presented together and never with any of the other options. The goal was to learn which option in each pair had a higher reward value. Then, in the transfer phase, all possible combinations of options were presented, and participants were asked to pick the option that they recalled having the higher value on each trial. The unique feature of this task is its manipulation of risk and outcome skew to fully dissociate the reference point, range adaptation, frequency encoding, and range-frequency models.

As shown in Table 3.1, the task manipulates Skew (negative or positive) and Maximizing Option (safe or risky) in a 2 × 2 within-subjects design. There are four learning contexts, each comprised of a safe option (*x* points with certainty) and a risky option (*y* points with p = .80; *z* points otherwise). The infrequent and relatively extreme outcomes from the risky option determine the skew of the contextual distribution. When the infrequent outcomes are *low* compared to the

most frequent outcomes, the distribution is negatively skewed (Contexts NR and NS). On the other hand, when the infrequent outcomes are *high*, the distribution is positively skewed (Contexts PR and PS). Importantly, the risky option produces a better outcome than the safe option 80% of the time in the negative skew contexts but only 20% of the time in the positive skew contexts. The second manipulated variable is the option that maximizes expected payoffs: In half of the contexts (NR and PR), the risky option has a higher EV while in the other half (NS and PS), the safe option has a higher EV. Note that the average point value of the four contexts increases from NR to NS to PR to PS, which serves to distinguish between absolute and context-dependent encoding.

Table 3.1 Summary of the Instrumental Learning Task in Experiment 2

Context	Lower EV option	Higher EV option
NR Skew: negative Maximizing Option: risky	NSL_{20} 20 (1.00)	NRH_{22} 25 (.80) 10 (.20)
NS Skew: negative Maximizing Option: safe	NRL_{24} 27 (.80) 12 (.20)	NSH_{26} 26 (1.00)
PR Skew: positive Maximizing Option: risky	PSL_{28} 28 (1.00)	PRH_{30} 27 (.80) 42 (.20)
PS Skew: positive Maximizing Option: safe	PRL_{32} 29 (.80) 44 (.20)	PSH_{34} 34 (1.00)

Note. Each option is named according to whether its local context has a negatively (N) or positively (P) skewed outcome distribution, whether it is safe (S) or risky (R), and whether it has the lower (L) or higher (H) expected value within its local context (expected values appear as subscripts). Each option is associated with one (safe) or two (risky) different outcomes, each occurring with a specific frequency (shown in parentheses as a relative frequency). EV = expected value.

Participants chose between the pairs of options in Table 3.1 during the learning phase, with each pair presented multiple times in an interleaved trial sequence. The design of this task allows for the identification of several different choice strategies just from looking at the preferred options in each pair. For example, a strategy of selecting actions that frequently produce better outcomes would favor NRH_{22}, NRL_{24}, PSL_{28}, and PSH_{34}, the options that produce a better outcome 80% of the time. General risk-seeking would lead to a preference for the risky options, while general risk-aversion would favor the safe options. The next section explains why the reference point and range adaptation models predict similar, EV-maximizing choice patterns in the learning phase, while the frequency encoding and range-frequency models predict a preference for the options that usually yield the best outcomes. It also shows that all four models make very distinct predictions for the subsequent transfer phase.

3.1 Model Predictions

The REFERENCE, RANGE, FREQUENCY, and RANGE-FREQUENCY models were simulated multiple times across a grid of parameter values in the task above. The parameter ranges were similar to those used in the simulations for Experiment 1 (REFERENCE: $\alpha_c, \alpha_u, \alpha_V \in \{.10, .15, \ldots, .50\}$, $\beta = .50$; RANGE: $\alpha_c, \alpha_u, \alpha_R \in \{.10, .15, \ldots, .50\}$, $\beta = 5$; FREQUENCY: $\alpha_c, \alpha_u \in \{.10, .15, \ldots, .50\}$, $w_F \in \{.50, .55, \ldots, .90\}$, $\beta = 5$; RANGE-FREQUENCY: $\alpha_c, \alpha_u, \alpha_R \in \{.1, .2, .3, .4, .5\}$, $w_F \in \{.3, .4, .5, .6, .7\}$, $\beta = 5$). Results were averaged across the various parameter configurations.

Simulation results for the REFERENCE model are shown in Figure 3.1. The reference points $V_t(s)$ are initialized to 27, the midpoint of the global reward distribution, and converge across the learning phase to the mean reward in each context (Figure 3.1A). Note that the rate of convergence could be increased or decreased by adjusting α_V. Simulated agents learn to prefer the EV-maximizing options in the learning phase (Figure 3.1B); however, the underlying Q values do not align with the objective values of the eight choice options. After the reference points stabilize at the mean of each context, the Q values of the higher-valued options in each context converge to 1.00, while the Q values of the lower-valued options converge to −1.00 (Figure 3.1C). In this task, there is a 2-point EV difference between the options in each context, which implies a 1-point difference between each option and the contextual average. The mean-centering mechanism leads to an overall preference in the transfer phase for options that are locally optimal during the learning phase, even when choosing these options violates EV-maximization (e.g., NRH_{22} is preferred over PRL_{32}; Figure 3.1D). Choice rates are close to chance when both options are locally optimal or suboptimal (e.g., NRL_{22} and PSH_{34}). It is important to note that the REFERENCE model predicts no effects of skew in either phase of the task.

Simulation results for the RANGE model are shown in Figure 3.2. The context-level variables $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are initialized to the global minimum (10) and maximum (44) rewards and gradually converge across learning trials to the local minimum and maximum in their corresponding contexts. As a result, the subjective ranges begin at 34 on the first trial and

eventually converge to 15, the range of outcomes in all four contexts (Figure 3.2A). The rate of convergence could be modulated by adjusting α_R. Simulated agents learn to choose the EV-maximizing options in each choice pair with increasing experience (Figure 3.2B).

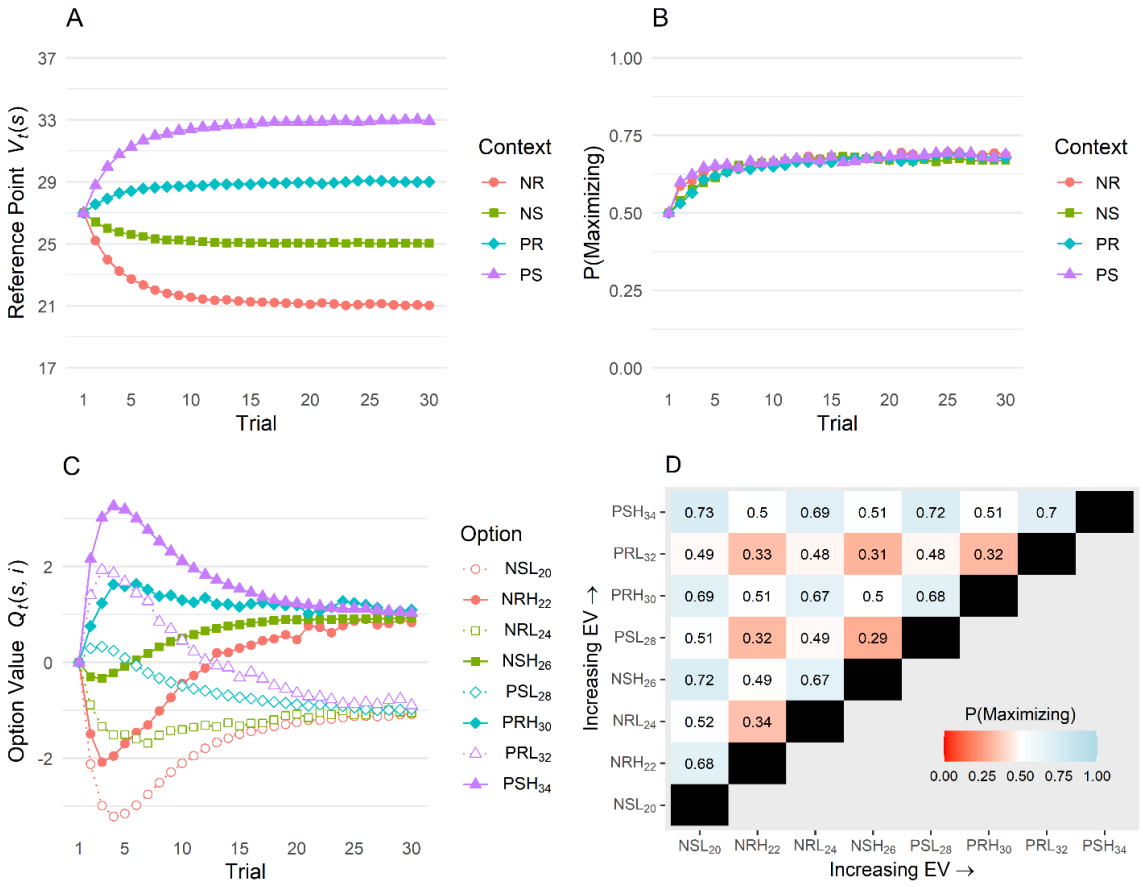


Figure 3.1 REFERENCE Model Simulations for Experiment 2. The REFERENCE model assumes that agents track a running estimate of the average reward in each context and evaluate options based on how their outcomes compare to the contextual average. (A) Learning of the reference points (i.e., average rewards) across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option Q values across the learning phase. (D) Pairwise preferences in the transfer phase. The numbers in each cell represent the proportion of times the EV-maximizing option (the row option) was selected.

In the first few trials of the learning phase, the Q values (initialized to 0.50) approximate the range values of the eight options with respect to the *global* reward distribution. This is because $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ are initialized to the global minimum and maximum rewards.

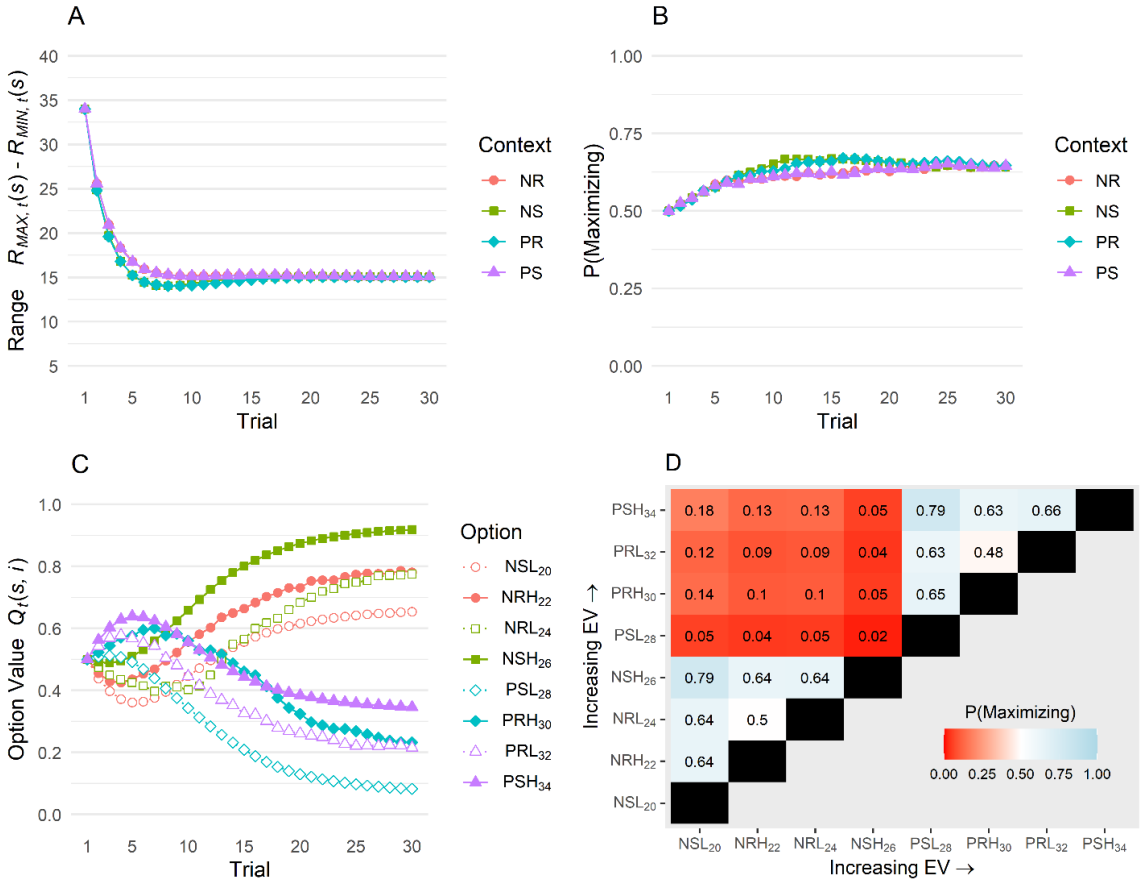


Figure 3.2 RANGE Model Simulations for Experiment 2. The RANGE model assumes that agents learn the smallest and largest rewards in each context and evaluate options based on where their outcomes fall along the contextual range. (A) Learning of the range of rewards across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option Q values across the learning phase. (D) Pairwise preferences in the transfer phase. The numbers in each cell represent the proportion of times the EV-maximizing option (the row option) was selected.

As $R_{MIN,t}(s)$ and $R_{MAX,t}(s)$ converge toward the endpoints of the four contexts, the Q values gradually adapt to the range of values in the *local* reward distributions. A side effect of this process is that it causes options in the negative skew contexts, where most outcomes are above the midpoint of the range, to finish with higher Q values than options in the positive skew contexts, where most outcomes are below the midpoint (Figure 3.2C). In the transfer phase, this leads to irrational preferences for the negative skew options over the positive skew options even though the former have lower EVs (Figure 3.2D). In summary, although the RANGE and REFERENCE models make similar predictions in the learning phase, they predict very different patterns in the transfer phase. The RANGE model predicts an overall attraction to options that come from contexts with negatively skewed reward distributions, whereas the REFERENCE model predicts no effect of skew.

Simulation results for the FREQUENCY model are shown in Figure 3.3. In this task, most of the outcomes in the negative skew contexts favor the risky option, whereas most of the outcomes in the positive skew contexts favor the safe option (see Table 3.1). Because frequency values are computed as a proportion of favorable outcome comparisons, the model predicts a preference for the risky options in the negative skew contexts but a reversed preference in the positive skew contexts. This leads to EV-maximizing behavior in Contexts NR and PS but suboptimal behavior in Contexts NS and PR (Figure 3.3A). The ordering of the Q values (initialized to 0.50) at the end of the learning phase reflects the influence of frequency information: Options that produce better

outcomes most of the time finish with higher Q values than their contextual counterparts (Figure 3.3B). In the transfer phase, violations of EV-maximization occur for choices between a lower-valued option that produced mostly high-ranking outcomes in its local context and a higher-valued option that produced mostly low-ranking outcomes in its local context (e.g., NRH_{22} is preferred over PRH_{30}; Figure 3.3C). Importantly, the FREQUENCY model predictions are very distinct from the predictions of the other two models in both the learning phase and transfer phase, allowing us to identify frequency encoding and distinguish it from reference-point centering and range adaptation.

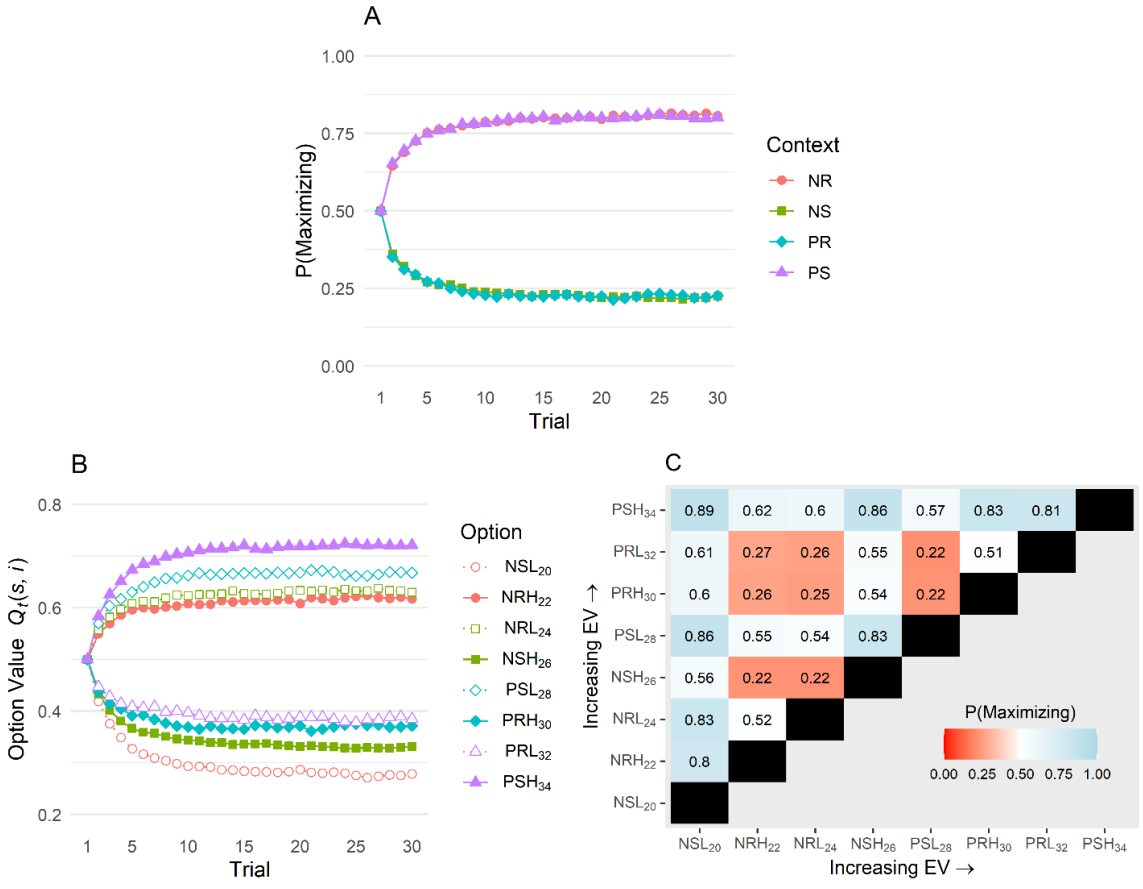


Figure 3.3 FREQUENCY Model Simulations for Experiment 2. The FREQUENCY model assumes that agents maintain exemplar representations of each context

and evaluate options based on the ranks of their outcomes within the contextual distribution. (A) Predicted probability of EV-maximizing choice across the learning phase. (B) Evolution of the option Q values across the learning phase. (C) Pairwise preferences in the transfer phase. The numbers in each cell represent the proportion of times the EV-maximizing option (the row option) was selected.

Finally, Figure 3.4 shows simulation results for the RANGE-FREQUENCY model. Its predictions are a compromise between the predictions of the previous two models. Like the RANGE model, it assumes that agents learn the minimum and maximum rewards in each context and evaluate outcomes based on where they fall along the range (Figure 3.4A). The model also borrows from the FREQUENCY model in implementing ordinal comparisons between the outcomes on each trial and recently experienced outcomes that are brought to mind. The model's simulated choice pattern in the learning phase resembles that of the FREQUENCY model (Figure 3.4B). However, the underlying Q values also reflect the range component, with options in negative skew contexts tending to have higher Q values than options in the positive skew contexts (Figure 3.4C). Although the range component has little to no effect in the learning phase, it has a noticeable effect in the transfer phase: Simulated agents show a robust preference for the options from negative skew contexts despite their lower objective values (Figure 3.4D). Other violations of EV-maximization in the transfer phase are driven by frequency information carried over from the learning phase (e.g., PSL_{28} is preferred over PRL_{32}).

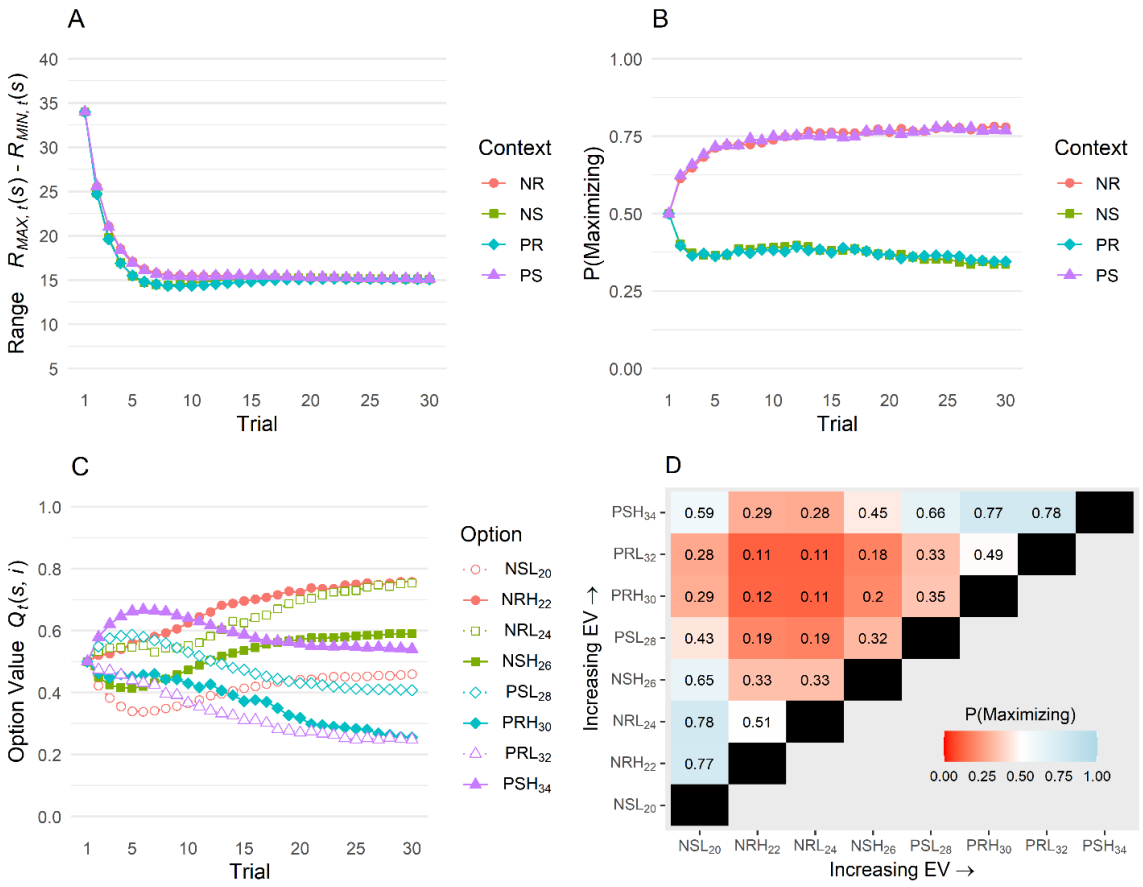


Figure 3.4 RANGE-FREQUENCY Model Simulations for Experiment 2. The RANGE-FREQUENCY model represents a compromise between range adaptation and frequency encoding models. (A) Learning of the range of rewards across the 30 learning phase trials for each context. (B) Predicted probability of EV-maximizing choice across the learning phase. (C) Evolution of the option *Q* values across the learning phase. (D) Pairwise preferences in the transfer phase. The numbers in each cell represent the proportion of times the EV-maximizing option (the row option) was selected.

3.2 Method

Our recruitment methods, experimental design, procedures, and data analysis plans were preregistered on the Open Science Framework (https://osf.io/xpn5g).

Participants

We used Prolific to recruit 50 participants (18 men, 29 women, 2 non-binary, 1 unstated; ages 18 – 64, *M* = 29.90, *SD* = 10.55) for an online experiment that was administered via Qualtrics. The inclusion criteria were the same as in Experiment 1. Sample size was based on a previous study that used a similar task design (Hayes & Wedell, in press). For control group participants in that study, the average proportion of EV-maximizing choices for transfer pairs in which context favored the objectively lower-valued option was .28 (*SD* = 0.29). To detect an effect of this magnitude with .90 power, 21 participants would be required (one-sample t-test comparing against chance, two-tailed, *d* = .76, α = .05). It took participants just over 33 minutes on average to complete the experiment. Participants were told that the points they earned in the task would be converted proportionately to real money and added to their participation payment, but they were not given the conversion rate (100 points = $0.04; mean bonus = $2.54). Participants provided informed consent and all aspects of this study were approved by the Institutional Review Board at the University of South Carolina.

Design

Learning Phase. The choice task in Experiment 2 was a two-part instrumental learning task with a learning phase and transfer phase. The learning phase employed a 2 × 2 within-subjects design with factors Skew (negative or positive) and Maximizing Option (safe or risky) (Table 3.1). There were four choice contexts formed by the combination of these two factors, each containing two options. A key aspect of this design is that the average point value of the

negative skew contexts (NR and NS) was lower than the average point value of the positive skew contexts (PR and PS), so that a consistent preference for the negative skew options in the transfer phase would be a strong indicator of range adaptation (see Figure 3.2D).

During the learning phase, the four contexts were each presented on 30 trials. Trial order was shuffled for each participant so that the contexts were randomly interleaved. Each choice option was always presented along with the other option in its context and both options appeared an equal number of times on the left and right side of the screen. Randomly generated identicons were used as option cues, and the assignment of cues to the eight options was randomized for each participant. The cues for the two options in each context had the same color (red, orange, green, or blue). In total, the learning phase contained 122 trials (4 contexts × 30 repetitions, plus 2 attention check trials).

Transfer Phase. The transfer phase consisted of choices between all possible pairs of options without feedback. With eight options, there were $\binom{8}{2}$ = 28 possible choice pairs and each was repeated four times. In total, the transfer phase contained 114 trials (28 choice pairs × 4 repetitions, plus 2 attention check trials). Trial order was shuffled for each participant and options appeared an equal number of times on the left and right side of the screen.

Procedure

Learning Phase. The instructions for the learning phase informed participants that they would be making repeated choices between two options at

a time to gain points. Each choice that they made would result in points, but participants were told that some options were more rewarding than others. The explicit goal was to gain as many points as possible. Participants were told that on each trial they would see the points produced by the chosen and nonchosen options, but only the chosen option's points counted toward their total (the running total was not visible).[6] They were also told that the experiment contained two parts and that both parts must be completed in one sitting. Participants were not given any specific details about the transfer phase, nor were they explicitly informed about the four contexts.

Each trial began with a 0.5 s fixation followed by the presentation of two option cues arranged horizontally on screen with the message, "Please make a choice" (Figure 3.5). Participants indicated their choice by clicking on one of the option cues. Following another 0.5 s fixation, participants received complete feedback on the outcomes from the chosen and nonchosen option, with the chosen cue indicated by a black border. The number of points produced by each option appeared just below the cues. Trials were self-paced.

Transfer Phase. The transfer phase instructions informed participants that they would once again be making binary choices, but some of the pairings may not be ones that they had previously seen. They were told that although

[6] Participants were shown an example of the choice feedback display with two options producing outcomes of 10 and 44 points. Because these outcomes correspond to the global minimum and maximum rewards, exposing participants to these outcomes in the instructions justified our choice of initial values for the RL models.

they would not see any points, the program would record the number of points they won from the chosen options and add it to their total. Finally, they were reminded of the goal of finishing with as many points as possible.

Each trial began with a 0.5 s fixation followed by the presentation of two option cues arranged horizontally on screen with the message, “Please make a choice” (Figure 3.5). A reminder message stating that points were being recorded appeared at the top of the screen. Trials were self-paced.

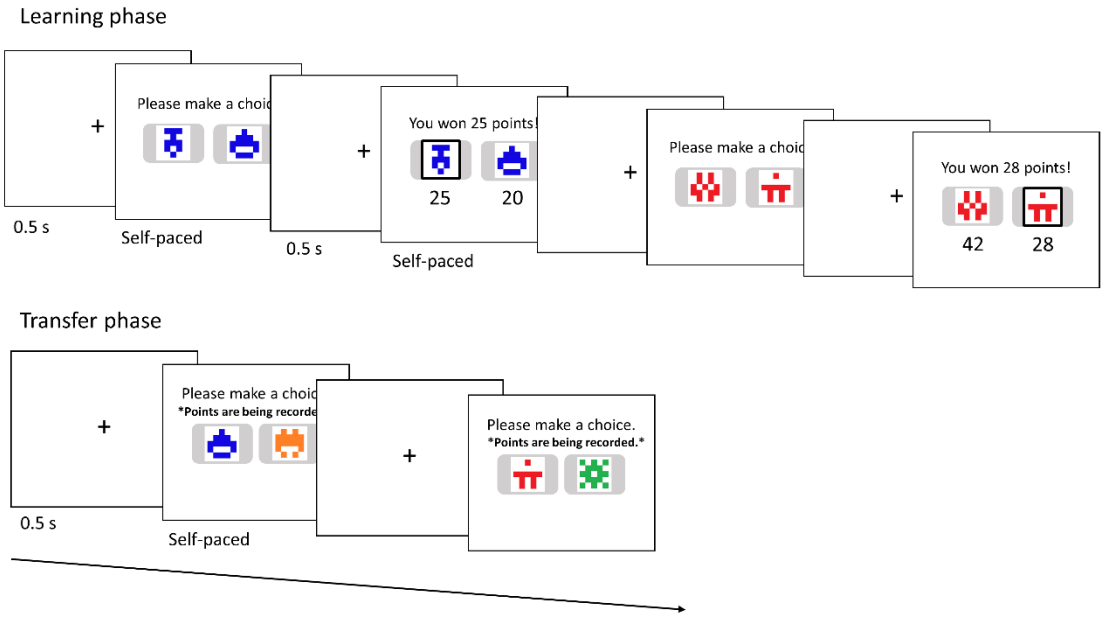


Figure 3.5 Trial Timeline for Experiment 2. Choice options were represented by cues (random identicons) that were the same color (red, orange, green, blue) for both options within a context. In the learning phase, each trial began with a prompt to choose between two options belonging to the same context. Each option appeared an equal number of times in the left and right position across the 30 trials. Following the participant’s choice and a 0.5 s fixation, complete feedback was presented from both options. The chosen option was indicated with a black border. Context presentations were randomly interleaved across trials. In the transfer phase, participants chose between all possible pairs of options without receiving feedback.

Attention Checks. There were four attention check trials at random points throughout the task, two in the learning phase and two in the transfer phase. The procedure for the attention checks was the same as in Experiment 1. Participants were excluded from the analysis if they failed more than one check.

Post-Task Questions. After completing the transfer phase, participants answered the same post-task questions as in the first experiment. First, they were shown the eight option cues in a randomized order and asked to arrange them from highest to lowest value. Second, participants were shown the two cues that belonged to each of the original learning contexts and asked to estimate the (1) average, (2) lowest, and (3) highest outcomes produced by the two options. Participants responded using a slider that ranged from 5 to 50 points. The order of context presentations was randomized.

Data Analysis and Modeling

For the learning phase data, a repeated-measures analysis of variance (ANOVA) was used to analyze the proportion of EV-maximizing choices as a function of Skew, Maximizing Option, and Block (five blocks of six trials per context). Based on our *ex-ante* model simulations, we expected to find a significant three-way interaction if the FREQUENCY or RANGE-FREQUENCY models are correct (see Figures 3.3A and 3.4B), whereas only a main effect of Block would be expected if the REFERENCE or RANGE models are correct (see Figures 3.1B and 3.2B).

For the transfer phase data, we analyzed choice rates for the eight options (i.e., the number of times an option was chosen, divided by the number of times it was presented; see Bavard et al., 2018) using a repeated-measures ANOVA with Skew (0 = positive, 1 = negative), Risk (0 = safe, 1 = risky), and Local Optimality (0 = no, 1 = yes) as factors. Local Optimality refers to whether the option was the maximizing or non-maximizing option in its learning context. According to the REFERENCE model, we should expect a significant main effect of Local Optimality but no effects of Skew or Risk. On the other hand, the RANGE model would produce a significant main effect of Skew. More complex patterns of effects would be expected if the FREQUENCY or RANGE-FREQUENCY models are correct.

For the model-based analysis, we fit the Q-learning model, the four context-dependent models, and two additional models to each participant's choice data using maximum likelihood methods. The models were compared based on their out-of-sample predictive accuracy in the transfer phase and BIC values. Out-of-sample prediction was carried out as follows. First, for each participant, the transfer phase data was randomly partitioned into four folds, with each fold containing seven different choice pairs. The out-of-sample prediction was performed in four iterations. Each iteration involved training the model on choices in the learning phase and three of the four transfer folds.[7] Then, the best-

[7] Our simulations revealed that training the models on at least a portion of the transfer choices was critical for ensuring that all parameters were identifiable. For example, the reference point learning rate (α_V) in the REFERENCE model cannot be identified from the learning phase data alone, given the design of our task.

fitting parameters were used to compute the out-of-sample log-likelihood of the choices in the remaining fold. The results were summed across the four folds and the model with the highest average out-of-sample log-likelihood was selected as the best model. We also tested each model's ability to generate the observed choice patterns when conditioned on the best-fitting parameters (Palminteri, Wyart, et al., 2017). Code for reproducing the analyses is available at https://osf.io/br3fq/.

3.3 Results

Learning Phase

Figure 3.6A shows the proportion of EV-maximizing choices across the 30 learning trials for each context. It is clear from the figure that the proportion of maximizing choices was highest overall in the NR and PS contexts—i.e., choice sets in which the maximizing option frequently produced better outcomes than the non-maximizing option. In addition, the rate of maximizing choices appeared to increase across trials in the positive skew contexts (PR and PS) while remaining relatively stable in the negative skew contexts (NR and NS).

The 30 trials for each context were divided into five blocks of six trials and the proportion of maximizing choices within blocks was computed for each participant. The choice proportions were then submitted to a 2 (Skew) × 2 (Maximizing Option) × 5 (Block) repeated-measures ANOVA. There was a main effect of Block, $F(4, 196) = 2.49$, $p = .044$, $\varepsilon = .86$, $\eta_p^2 = .05$, and a significant Skew × Block interaction, $F(4, 196) = 5.18$, $p < .001$, $\varepsilon = .85$, $\eta_p^2 = .10$.

Maximization rates changed across blocks in the positive skew contexts with a significant positive linear trend (contrast coefficient = 0.29), $t(49) = 4.24$, $p < .001$, and negative quadratic trend (contrast coefficient = −0.24), $t(49) = 2.93$, $p = .005$. On the other hand, none of the trends were significant in the negative skew contexts ($ps > .33$). The most critical effect was the Skew × Maximizing Option interaction, $F(1, 49) = 29.19$, $p < .001$, $\eta_p^2 = .37$. In the negative skew contexts, maximization rates were higher when the maximizing option was risky compared to when it was safe, $t(49) = 5.35$, $p < .001$, while the reverse was true in the positive skew contexts, $t(49) = 3.90$, $p < .001$. Taken together, the rate of optimal choices was higher in those contexts where the maximizing option frequently produced the best outcomes (NR context: $M = .63$, 95% CI = [.59, .68]; PS context: $M = .64$, 95% CI = [.58, .70]) compared to those contexts where the non-maximizing option frequently produced the best outcomes (NS context: $M = .42$, 95% CI = [.37, .47]; PR context: $M = .51$, 95% CI = [.45, .56]). This result is consistent with the FREQUENCY and RANGE-FREQUENCY models. All other effects were nonsignificant ($ps > .05$).

Transfer Phase

In the transfer phase, participants encountered all possible pairs of options and were tasked with choosing the higher valued option on each trial. Feedback was not presented, and each of the 28 choice pairs was repeated four times. Figure 3.6B shows the pattern of transfer preferences averaged across participants, with each cell showing the mean percentage of times that the option on the vertical axis (“Option 2”) was chosen over the option on the horizontal axis

("Option 1"). Higher numbers represent greater payoff maximization since the EV of Option 2 is always greater than the EV of Option 1.

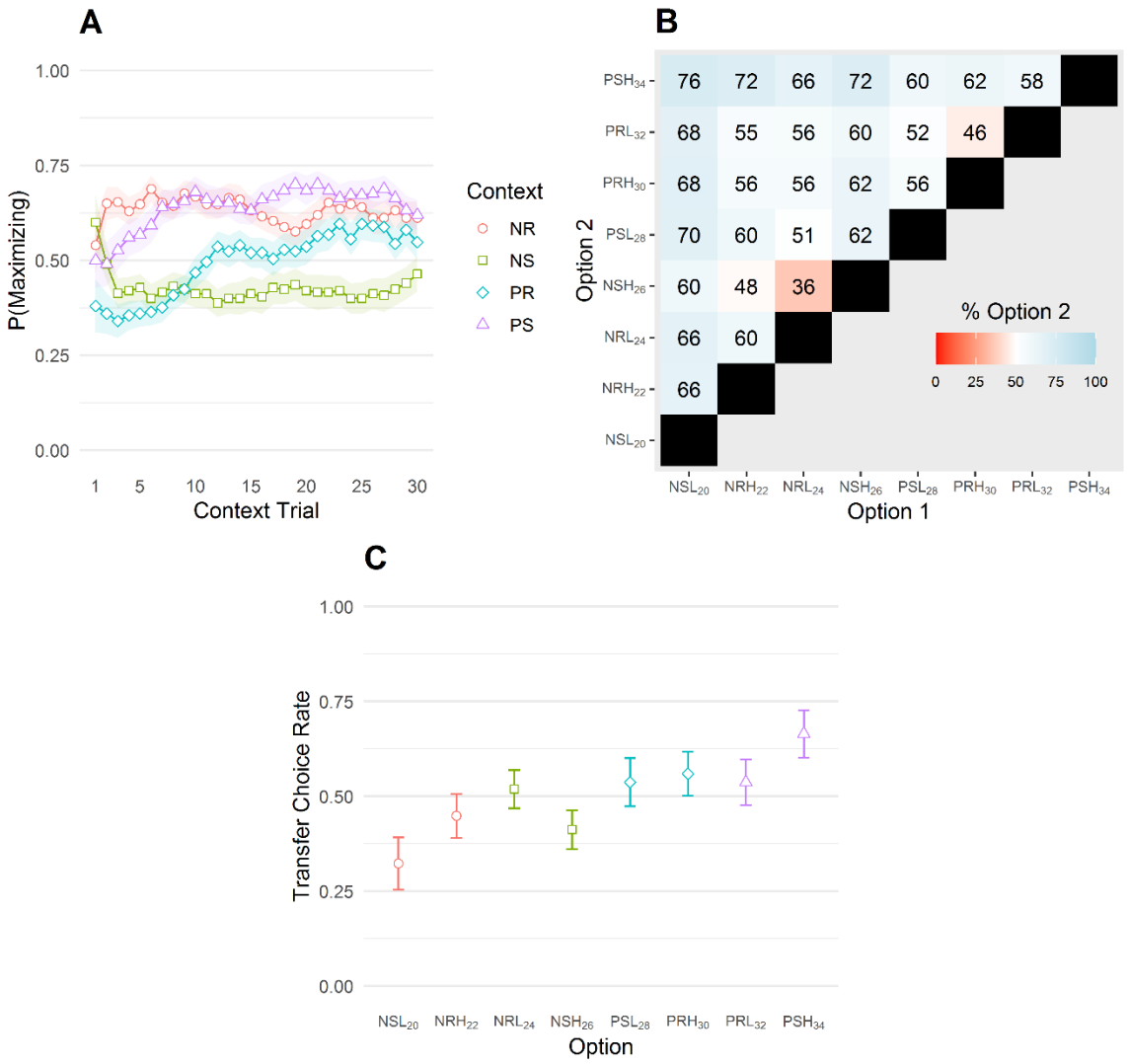


Figure 3.6 Learning and Transfer Phase Results for Experiment 2. (A) Mean proportion of EV-maximizing choices across the 30 learning trials for each context. Choices were smoothed at the individual level using a 5-trial rolling average prior to averaging across individuals. Error bands represent +/− 1 standard error. (B) Pairwise choice preferences in the transfer phase. Each cell shows the mean percentage of times the EV-maximizing option (Option 2) was selected, averaging across individuals. (C) Mean choice rates for each option in the transfer phase, defined as the number of times an option was selected divided by the number of times it was presented. Error bars represent 95% confidence intervals.

We correlated the pattern of pairwise preferences in Figure 3.6B with the patterns that were simulated *ex-ante* by the four context-dependent RL models (see Figures 3.1D, 3.2D, 3.3C, and 3.4D). Note that the simulations were generated using a range of different parameter values and prior to seeing the data. The correlation was strongest for the FREQUENCY model (r = 0.72), followed by the RANGE-FREQUENCY model (r = 0.24) and the REFERENCE model (r = 0.15). The correlation between the empirical data and the RANGE model was negative (r = −0.34). These results provide initial support for the FREQUENCY model, but the fact that most of the maximization rates in Figure 3.6B are above 50% suggests that absolute encoding models may offer a viable account of the transfer phase data as well. Indeed, after simulating the Q-learning model over a range of parameter values, we found a considerable correlation between its predicted pattern and the observed pattern (r = 0.56).

To analyze the transfer phase, we computed choice rates for the eight options separately for each participant (i.e., the number of times an option was chosen, divided by the number of times it was presented; Bavard et al., 2018). The mean choice rates are shown in Figure 3.6C. Choice rates were submitted to a 2 (Skew) × 2 (Risk) × 2 (Local Optimality) repeated-measures ANOVA, where Risk refers to whether the option was safe or risky and Local Optimality refers to whether the option was locally optimal in its original encoding context. The results indicated a significant main effect of Skew, $F(1, 49) = 25.47$, $p < .001$, $\eta_p^2 = .34$, with options from positive skew contexts (M = .57, 95% CI = [.55, .60]) chosen more often than options from negative skew contexts (M = .43, 95% CI =

[.40, .46]). This is consistent with absolute value encoding since the positive skew contexts had larger rewards in our choice task, but it is inconsistent with the RANGE model, which predicted an overall preference for the options from negative skew contexts in our *ex-ante* simulations (see Figure 3.2D). There was also a significant main effect of Risk, $F(1, 49) = 4.12$, $p = .048$, $\eta_p^2 = .08$, and a significant Skew × Risk interaction, $F(1, 49) = 11.91$, $p = .001$, $\eta_p^2 = .20$. The risky options from negative skew contexts (NRH_{22} and NRL_{24}) were chosen more often than the safe options from negative skew contexts (NSL_{20} and NSH_{26}), $t(49) = 3.99$, $p < .001$, whereas the safe options from positive skew contexts (PSL_{28} and PSH_{34}) were chosen more often than the risky options from positive skew contexts (PRH_{30} and PRL_{32}), $t(49) = 1.81$, $p = .08$. This result represents a general preference for options that typically produced better outcomes in their local encoding contexts, which is the key behavioral signature of models that incorporate frequency encoding (FREQUENCY and RANGE-FREQUENCY). Finally, there was a significant main effect of Local Optimality, $F(1, 49) = 5.08$, $p = .029$, $\eta_p^2 = .09$, and a significant Risk × Local Optimality interaction, $F(1, 49) = 6.89$, $p = .012$, $\eta_p^2 = .12$, due to the safe options (but not the risky options) being chosen significantly more often when they were locally optimal compared to when they were locally suboptimal in the learning phase. However, these effects are not of particular interest for distinguishing between models. The Skew × Local Optimality interaction ($p = .06$) and the three-way interaction ($p = .55$) were both nonsignificant.

Model Comparison

We fit the Q-learning model, the four context-dependent encoding models, and two additional models to each participant's data using maximum likelihood methods. The two additional models both assume absolute value encoding but differ in how they update reward expectations in response to feedback (i.e., Equation 1). The first is a model that uses separate learning rates for positive and negative reward prediction errors (RPEs). When the learning rate for positive RPEs exceeds the learning rate for negative RPEs, decision makers become risk-seeking, whereas the opposite pattern leads to risk aversion (Niv et al., 2012). Thus, the RISK-SENSITIVE model includes three parameters (β, α_+, α_-). The second model uses separate learning rates for outcomes that confirm the decision maker's choice (i.e., positive RPEs for the chosen option and negative RPEs for the unchosen option) and outcomes that disconfirm their choice (i.e., negative RPEs for the chosen option and positive RPEs for the unchosen option). When the learning rate for confirmatory outcomes exceeds the learning rate for disconfirmatory outcomes, decision makers tend to repeat previous choices regardless of their consequences (Palminteri, Lefebvre, et al., 2017). This CONFIRMATION BIAS model also includes three parameters (β, α_{CON}, α_{DIS}). The RISK-SENSITIVE model was relevant for our task given the manipulation of risk level; however, the CONFIRMATION BIAS model has been shown to provide a better account of choice behavior in similar tasks (Palminteri, Lefebvre, et al., 2017).

The results of the relative model comparison are shown in Table 3.2. This time, the RANGE-FREQUENCY model had the highest mean out-of-sample log-

likelihood and lowest mean BIC, indicating that it was the best model overall. Paired *t*-tests revealed that the RANGE-FREQUENCY model was significantly better than the Q-learning, REFERENCE, RANGE, and RISK-SENSITIVE models according to both metrics. However, it was not significantly better than the simpler FREQUENCY model, and it was only marginally favored over the CONFIRMATION BIAS model based on out-of-sample log-likelihoods. Overall, these results support the frequency encoding hypothesis, but further tests were needed to establish a clear advantage of the frequency encoding models over the CONFIRMATION BIAS model. Note that the mean estimates of w_F from the FREQUENCY (.38) and RANGE-FREQUENCY models (.31) were lower than the values used in the *ex-ante* simulations (see Table 3.3 for a full summary of the parameter estimates).

Table 3.2 Model Comparison Results in Experiment 2

Model	Parameters	Out-of-sample log-likelihood (Transfer phase only)	BIC (Both phases)
Q-learning	3	−70.44*	309.03**
REFERENCE	4	−70.58*	309.01**
RANGE	4	−69.50*	307.61**
FREQUENCY	4	−68.41	298.20
RANGE-FREQ.	5	**−66.71**	**297.22**
RISK-SENSITIVE	3	−70.90**	309.96**
CONFIRM. BIAS	3	−70.53†	299.62

Note. Mean out-of-sample log-likelihood and Bayesian information criterion (BIC) values. The best model according to each metric is shown in bold. Significance tests reflect comparisons of each model to the best model using paired t-tests (df = 49). BIC = −2 × *LL* + *k* × ln(*n*), where LL is the maximized log-likelihood, *k* is the number of model parameters, and *n* is the number of observations.
†*p* < .10, **p* < .05, ***p* < .01.

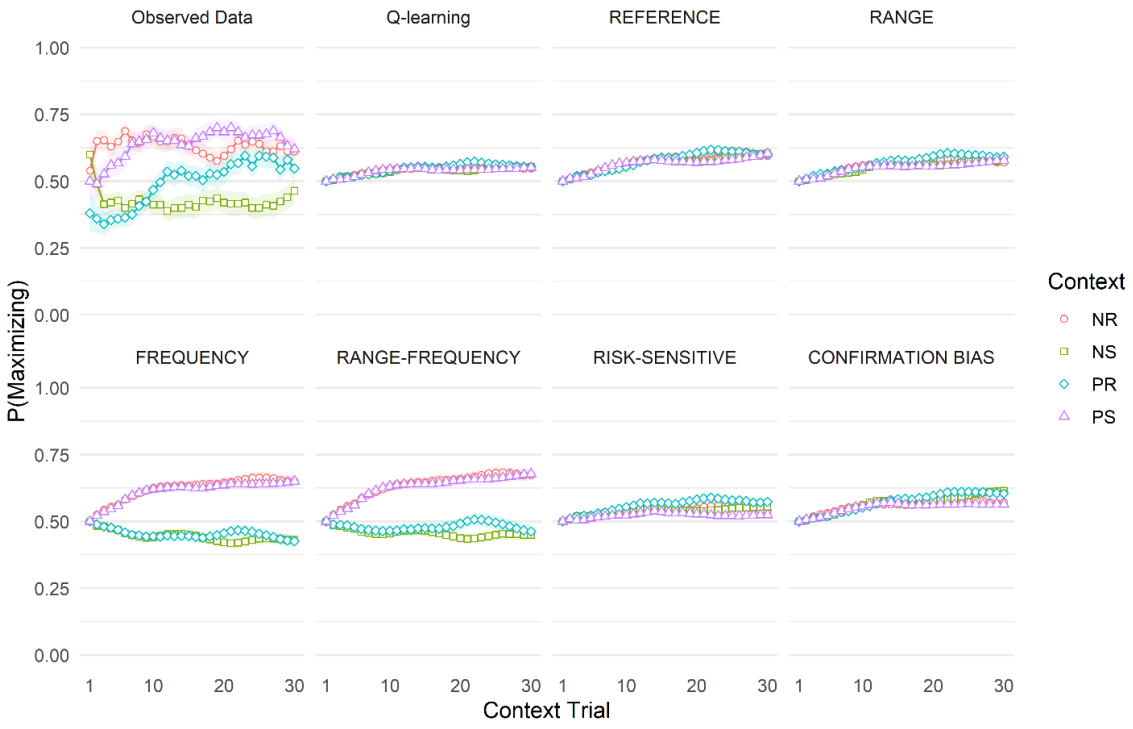
Table 3.3 Mean Parameter Estimates in Experiment 2

Model	β	α_c	α_u	α_V	α_R	w_F	α_+	α_-	α_{CON}	α_{DIS}
Q-learning	1.81 (5.46)	.37 (.39)	.36 (.39)	--	--	--	--	--	--	--
REFERENCE	3.21 (6.91)	.26 (.33)	.26 (.35)	.27 (.35)	--	--	--	--	--	--
RANGE	8.30 (7.58)	.32 (.38)	.27 (.36)	--	.21 (.36)	--	--	--	--	--
FREQUENCY	7.39 (7.32)	.30 (.34)	.29 (.32)	--	--	.38 (.37)	--	--	--	--
RANGE-FREQ.	9.24 (7.99)	.26 (.33)	.21 (.29)	--	.29 (.40)	.31 (.31)	--	--	--	--
RISK-SENSITIVE	1.84 (5.34)	--	--	--	--	--	.43 (.40)	.37 (.38)	--	--
CONFIRM. BIAS	1.46 (4.78)	--	--	--	--	--	--	--	.40 (.33)	.23 (.34)

Note. Standard deviations shown in parentheses. β = inverse temperature, α_c = chosen learning rate, α_u = unchosen learning rate, α_V = reference point learning rate, α_R = range learning rate, w_F = frequency value weighting, α_+ = learning rate for positive prediction errors, α_- = learning rate for negative prediction errors, α_{CON} = learning rate for confirmatory outcomes, α_{DIS} = learning rate for disconfirmatory outcomes.

Generative performance was assessed by simulating each model 100 times in the task using each participant's optimized parameters and averaging the predicted choice probabilities over iterations. Importantly, the models were not provided with participants' actual choice histories for the simulations. As shown in Figure 3.7, the simulated learning curves from the FREQUENCY and RANGE-FREQUENCY models most closely resembled the observed data. Indeed, the FREQUENCY and RANGE-FREQUENCY models were the only ones that accurately reproduced the critical Skew × Maximizing Option interaction in the learning phase, which we verified by conducting the same

ANOVA on the simulated datasets that was conducted on the empirical data. Both models favored the options that frequently produced the best outcomes in their local encoding contexts, resulting in significantly greater maximization rates in the NR and PS contexts compared to the NS and PR contexts. The REFERENCE and RANGE models failed to reproduce this interaction altogether, while the RISK-SENSITIVE and CONFIRMATION BIAS models produced significant Skew × Maximizing Option interactions but with an incorrect ordering of the mean choice proportions across contexts. In other words, none of the other models generated the key preference for options with higher frequency values.[8]



[8] It should also be noted that none of the models—including the FREQUENCY and RANGE-FREQUENCY models—was able to reproduce the significant Skew × Block interaction that was observed in the empirical data (i.e., increasing maximization rates in the positive skew contexts, but not in the negative skew contexts).

Figure 3.7 Learning Phase Choice Behavior in Experiment 2: Empirical Data vs. Model Simulations. Mean proportion of EV-maximizing choices across the 30 learning trials for each context. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. In all panels, choices were smoothed using a 5-trial rolling average prior to averaging across individuals. Error bands represent +/− 1 standard error.

Figure 3.8 shows the simulated patterns of pairwise choice preferences in the transfer phase. The observed data are shown in the upper left corner, and the correlations between the models and the observed data are displayed in the remaining panels. When conditioned on participants' best-fitting parameters, the FREQUENCY (r = 0.86) and RANGE-FREQUENCY models (r = 0.78) generated transfer preferences that were more strongly correlated with the observed preferences compared to the other models. The next highest correlation was attributed to the Q-learning model (r = 0.62) and the weakest correlation was attributed to the RANGE model (r = 0.40).

Figure 3.9 shows the observed and model-simulated choice rates for the eight options in the transfer phase. Most of the models generated patterns that were clearly inconsistent with the empirical data even after conditioning on the optimized parameters. For example, the Q-learning model generated choice rates that increased monotonically from the lowest- to the highest-valued option due to its reliance on absolute outcome encoding (the same was true for the RISK-SENSITIVE and CONFIRMATION BIAS models). The REFERENCE model generated a sawtooth-like pattern in which the locally optimal options in each

learning context were chosen more frequently than the locally suboptimal options. In contrast, participants chose the lower-valued option from the NS context (NRL_{24}) more often than the higher-valued option (NSH_{26}).

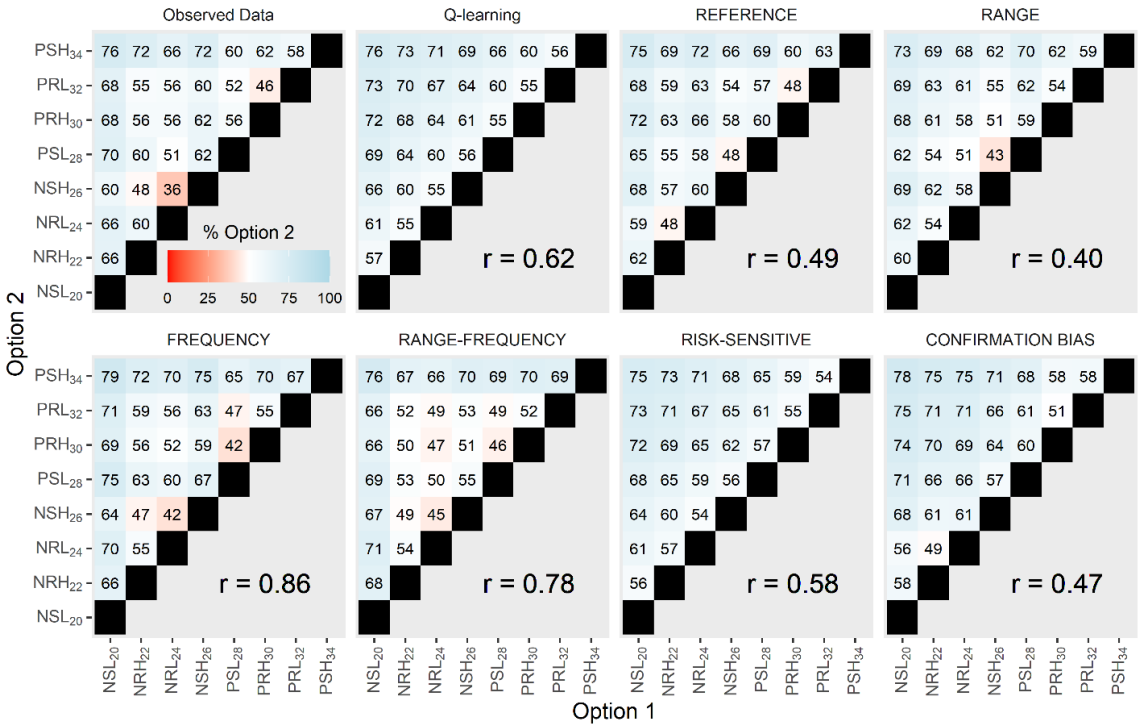


Figure 3.8 Transfer Phase Pairwise Choice Preferences in Experiment 2: Empirical Data vs. Model Simulations. Each cell shows the mean percentage of times the EV-maximizing option (Option 2) was selected, averaging across individuals. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. Also shown are the correlations between the empirical choice pattern and the patterns for each model.

The RANGE model predicted that the highest-valued option from the negative skew contexts (NSH_{26}) would be chosen more often than the lowest-valued option from the positive skew contexts (PSL_{28}), but participants exhibited the

opposite effect. Only the FREQUENCY and RANGE-FREQUENCY models were able to generate the critical Skew × Risk interaction, in which the options that frequently produced the best outcomes in their local encoding contexts (NRH_{22}, NRL_{24}, PSL_{28}, and PSH_{34}) were selected more often than the other options (tested using the same ANOVA that was conducted on the empirical choice rates). These results, when combined with the simulation results in the learning phase, demonstrate a clear advantage of the FREQUENCY and RANGE-FREQUENCY models over the other candidate models.

Figure 3.9 Transfer Phase Choice Rates in Experiment 2: Empirical Data vs. Model Simulations. Mean choice rates for each option in the transfer phase, averaged across individuals. Choice rate is defined as the number of times an option was selected divided by the number of times it was presented. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models

were not provided with participants' actual choices for the simulations. Error bars represent 95% confidence intervals.

Post-Task Questions

At the end of the experiment, participants were shown the eight option cues in a randomized order and asked to rank them from highest (1) to lowest (8) value. The mean reported ranks for each option are shown in Figure 3.10A. There was an increasing trend in ranks across the eight options, suggesting that participants were somewhat sensitive to the absolute values of options. In the contexts where the maximizing option frequently produced better outcomes (NR and PS), the maximizing options were given higher ranks than the non-maximizing options on average [NR: $t(49) = 3.53$, $p < .001$; PS: $t(49) = 2.20$, $p = .032$]. In contrast, there was no significant difference between reported ranks in the contexts where the non-maximizing option frequently produced better outcomes [NS: $t(49) = 1.10$, $p = .28$; PR: $t(49) = 0$, $p = 1.00$]. This suggests that the experienced frequency of favorable outcomes at least partially affected subjective judgments of the relative ranks of choice options.

Participants were also asked to estimate the (1) average, (2) lowest, and (3) highest outcomes in each context. The individual-level estimates and group averages are shown in Figure 3.10B−D. As in the first experiment, the individual estimates were noisy and the aggregate estimates often deviated significantly from the actual values of the outcome statistics. Participants seemed to be less accurate when estimating the minimum and maximum outcomes compared to the average outcomes in this experiment. Overall, participants appeared to be

somewhat sensitive to the relative differences between the outcome statistics of the four contexts.

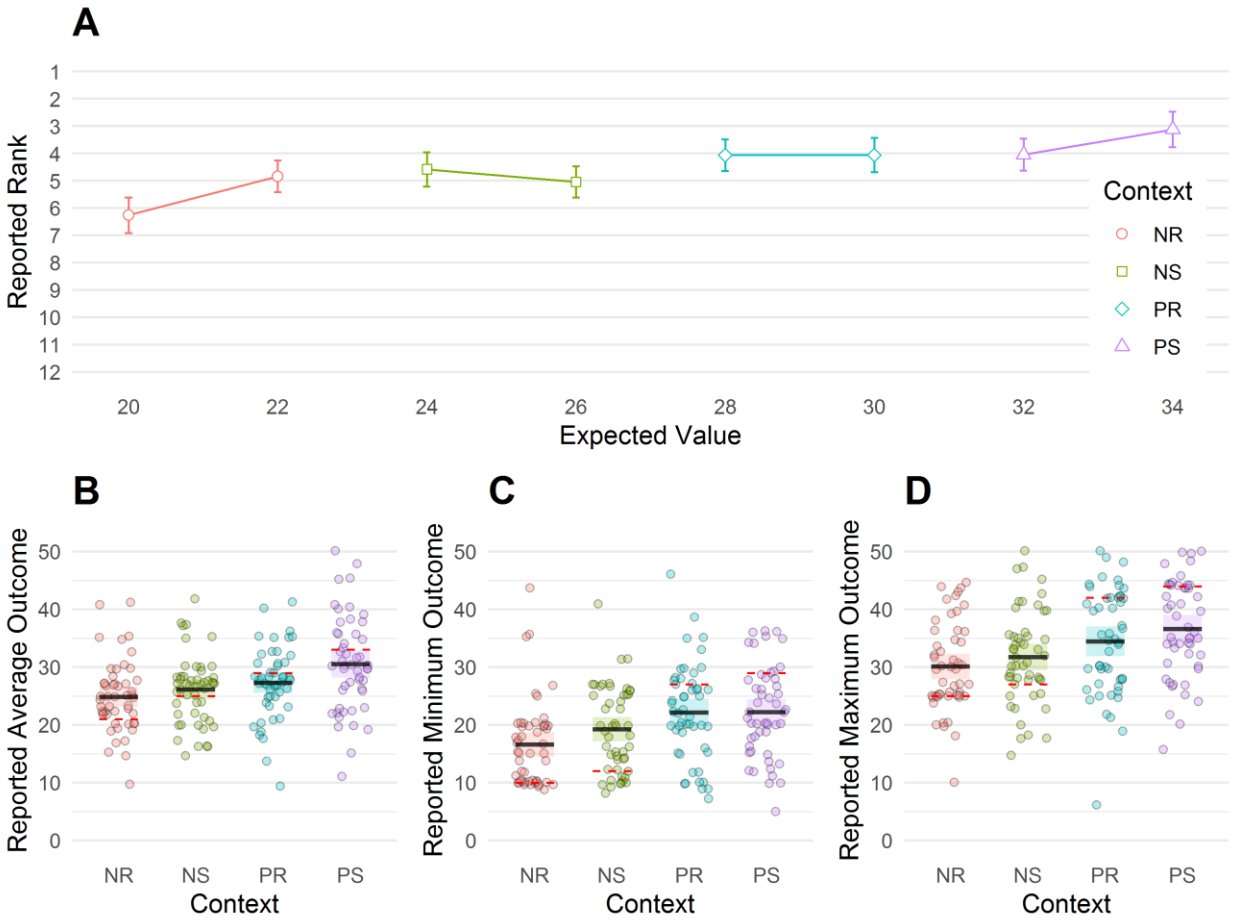


Figure 3.10 Responses to the Post-Task Questions in Experiment 2. (A) Mean reported ranks for each choice option. Error bars show 95% confidence intervals. Panels B-D show the reported average (B), minimum (C), and maximum outcomes (D) for each context. Individual estimates are shown as points and the group means are shown as solid black lines. Shaded boxes represent 95% confidence intervals. The true average, minimum, and maximum outcomes are shown as dashed red lines.

Separate ANOVAs were run on the estimated average, minimum, and maximum outcomes with Context as a repeated measures factor. These analyses were not preregistered and should be considered exploratory. The

effect of Context was significant in all three cases [average: $F(3, 147) = 8.33$, $p < .001$, $\varepsilon = .92$, $\eta_p^2 = .15$; minimum: $F(3, 147) = 10.33$, $p < .001$, $\varepsilon = .95$, $\eta_p^2 = .17$; maximum: $F(3, 147) = 8.55$, $p < .001$, $\varepsilon = .97$, $\eta_p^2 = .15$]. Estimates of the average outcomes showed a positive linear trend across contexts (contrast coefficient = 18.28), $t(49) = 4.29$, $p < .001$, but the quadratic and cubic trends were nonsignificant (ps > .20) (Figure 3.10B). This matches the actual pattern of average outcomes, which increased in a linear fashion across the four contexts (NR: 21, NS: 25, PR: 29, PS: 33). Estimates of the minimum outcomes showed a positive linear trend across contexts (contrast coefficient = 19.86), $t(49) = 4.84$, $p < .001$, with nonsignificant quadratic and cubic trends (ps > .11) (Figure 3.10C). This does not match the actual set of minimum outcomes, as the difference between the negative skew and positive skew contexts was much larger than the differences within each of the skew conditions, creating a step-like pattern (NR: 10, NS: 12, PR: 27, PS: 29). Similarly, estimates of the maximum outcomes increased linearly across contexts (contrast coefficient = 22.32), $t(49) = 4.98$, $p < .001$, but the quadratic and cubic trends were nonsignificant (ps > .72) (Figure 3.10D). This pattern again fails to capture the step-like pattern in the actual set of maximum outcomes (NR: 25, NS: 27, PR: 42, PS: 44).

3.4 Discussion

The purpose of Experiment 2 was to test competing theories of context-dependent RL using a choice task for which these theories make distinct predictions with regard to the effects of skewed outcome distributions. As in the first experiment, the results were most consistent with models that incorporate

frequency or rank-based encoding. A key behavioral signature of these models is a preference for options that frequently produce the best outcomes in their local encoding contexts. Model-free analyses confirmed the presence of this effect in both phases of the experiment. In the learning phase, the rate of payoff-maximizing choices was higher in contexts where the locally optimal option frequently produced the best outcomes. In the transfer phase, there was still a significant, albeit weaker, tendency to select these same options.

Model comparison indicated that the RANGE-FREQUENCY model provided the best overall explanation of the data. It was significantly favored over the Q-learning model, which assumes absolute encoding, as well as the REFERENCE and RANGE models, which implement context dependence via reference point centering or range adaptation, respectively. The RANGE-FREQUENCY and FREQUENCY models were the only ones that generated the key behavioral signature discussed above, i.e., a significant preference for options that frequently produced the best outcomes in their local contexts. The other models were unable to reproduce this effect. Further, the results could not be adequately explained by absolute encoding models that use separate learning rates for positive versus negative prediction errors (RISK-SENSITIVE model) or for outcomes that confirm versus disconfirm the decision maker's choice (CONFIRMATION BIAS model). The CONFIRMATION BIAS model was not significantly worse than the RANGE-FREQUENCY model according to the relative comparison criteria, but it failed to generate the key behavioral effects using participants' fitted parameters alone. This shows that its ability to make

accurate one-step-ahead predictions derived heavily from capitalizing on the temporal autocorrelation detected in past choices (Palminteri, Wyart, et al., 2017). Thus, the CONFIRMATION BIAS model is not well-suited as an explanatory model of the cognitive processes guiding context-dependent choice behavior. However, incorporating confirmation bias in the updating of expected rewards may further improve the performance of context-dependent RL models.[9]

Participants' subjective rankings of the options at the end of the experiment were mostly consistent with the aggregate choice rates in the transfer phase (cf. Figure 3.6C and 3.10A). Unlike in the first experiment, estimates of the lowest and highest outcomes in each context were less accurate than estimates of the average outcomes. In particular, participants were somewhat insensitive to the large difference between the endpoints of the negative skew contexts on the one hand, and the positive skew contexts on the other. However, the ordering of the most extreme outcomes across contexts was correct at the aggregate level.

[9] For example, we found that by swapping out the learning rates for chosen (α_c) and unchosen options (α_u) with learning rates for confirmatory (α_{CON}) and disconfirmatory outcomes (α_{DIS}), we could further improve the fit of the winning RANGE-FREQUENCY model in Experiment 2 (results not reported here).

CHAPTER 4

GENERAL DISCUSSION

There is ample evidence across multiple studies that subjective values are context dependent (for reviews, see Hunter & Daw, 2021; Rangel & Clithero, 2012; Seymour & McClure, 2008). This means that a given reward will be evaluated differently depending on the values of other rewards in the relevant context. As an illustration, a medium-sized reward may have a higher subjective value in the context of smaller rewards, but a lower subjective value in the context of larger rewards. This type of relative valuation has also been demonstrated in reinforcement learning (RL) tasks, where decision makers learn from the outcomes of previous choices to make optimal future choices (for a review, see Palminteri & Lebreton, 2021). For example, relative valuation aids in learning to avoid options that are worse than other options in the local environment, such as an option that frequently produces neutral outcomes in a gain context, and learning to prefer similar options if they are better than other options in the local environment, such as an option that frequently produces neutral outcomes in loss context (Palminteri et al., 2015). At the same time, a potential consequence of relative valuation is that individuals may prefer a favorable option from a low-value context over an unfavorable option from a high-value context, even if the latter has a higher expected payoff (Bavard et al., 2018;

Hayes & Wedell, in press). Thus, context-dependent valuation in RL can have both adaptive and maladaptive consequences (Bavard et al., 2021).

The aim of this dissertation was to test between competing computational mechanisms responsible for these effects. There are many theories that could potentially explain context dependence in RL. One possibility is that decision makers compare the outcomes of each choice to a contextual reference point—namely, the average outcome in the local context—so that options with better-than-average payoffs acquire a positive subjective value and options with worse-than-average payoffs acquire a negative subjective value. This mechanism was employed by the REFERENCE model, which dynamically tracks an estimate of the average outcome in each decision context (Palminteri et al., 2015; Palminteri & Lebreton, 2021). A second possibility is that decision makers encode where the outcomes of each choice fall with respect to the range of outcomes in the local context, so that options with payoffs near the top of the range acquire a higher subjective value than options with payoffs near the bottom of the range. This mechanism was employed by the RANGE model, which tracks a dynamic estimate of the minimum and maximum outcomes in each decision context (Bavard et al., 2021; Palminteri & Lebreton, 2021). A third possibility is that decision makers rank the outcomes of each choice within a sample of recently experienced outcomes from the local context (e.g., Stewart et al., 2006), so that options with high-ranking outcomes acquire a higher subjective value than options with low-ranking outcomes. This mechanism was implemented by the FREQUENCY model, which unlike the REFERENCE and RANGE models,

assumes an exemplar-based representation of the local context. Finally, a fourth possibility is that subjective values are based on a weighted combination of the range and frequency mechanisms (Parducci, 1965, 1995), which was the motivation behind the RANGE-FREQUENCY model. Previous studies have found support for the reference point (Palminteri et al., 2015) and range adaptation mechanisms (Bavard et al., 2018, 2021); however, to the best of our knowledge, no studies have directly tested RL models based on the frequency or range-frequency mechanisms. Further, range adaptation and frequency encoding were confounded in the choice tasks used in prior studies (see Table 1 for an example), making it difficult to determine which mechanism offered the best explanation of behavior.

We conducted two fully incentivized online experiments to test the REFERENCE, RANGE, FREQUENCY, and RANGE-FREQUENCY models using choice tasks for which the models make distinct predictions. In Experiment 1, we modified a task that had been previously used to disentangle competing theories of context effects in price perception (Niedrich et al., 2001) so that it could be leveraged to study context effects in RL. There were three “target” options in our task that produced the same outcomes and had the same expected payoffs but belonged to separate choice sets (i.e., contexts). One of the target options belonged to a negatively skewed context in which its outcomes were near the top of the range but had lower ranks. The second target option belonged to a positively skewed context in which its outcomes were near the bottom of the range but had higher ranks. The expected payoffs of these target options were

equal to the average rewards in their respective contexts. Finally, the third target option belonged to a positively skewed context in which its outcomes were near the top of the range *and* had higher ranks. The expected payoff of this target was well above the average reward in its context. The task involved an initial learning phase, in which participants made repeated choices within the separate contexts (presented in random order) and received complete feedback, followed by a transfer phase, in which they chose between options from different contexts without receiving feedback. The most diagnostic transfer choices were those that involved two of the target options, as the models predicted different preference relations among the targets depending on how their outcomes compared to other outcomes in their respective contexts (see Table 3). The results showed a significant aggregate preference for the targets from the positive skew contexts over the target from the negative skew context. This finding strongly supports the frequency encoding mechanism, since the only advantage of the positive skew targets over the negative skew target is that their outcomes had higher ranks within the local contextual distribution. Quantitative model comparison confirmed that the FREQUENCY model was superior to the other models, and we demonstrated that it was able to reproduce the aggregate choice preferences in the transfer phase using the parameter estimates from our sample. The RANGE-FREQUENCY model was also a viable candidate, but its extra parameter was not necessary to capture the empirical choice patterns.

Experiment 2 introduced a choice task that manipulated both outcome skew and risk. There were four separate choice contexts during the learning

phase, each consisting of a safe option that produced a certain outcome, and a risky option that produced one outcome with 80% probability and another with 20% probability. Whether the more probable outcome was high (negative skew) or low (positive skew) varied across contexts. We also varied which option (risky or safe) was locally optimal to avoid a confound between payoff maximization and risk preference. An important feature of the task is that the risky options produced better outcomes more often in the negative skew contexts, whereas the safe options produced better outcomes more often in the positive skew contexts. In the learning phase, participants exhibited a robust preference for the options that frequently produced the best outcomes in their local contexts. This tendency to select options with higher ranking outcomes was dissociable from risk preference and payoff maximization due to the structure of the task. Most importantly, it was a key behavioral signature of frequency encoding. In the subsequent transfer phase (choices between all possible pairs of options without feedback), participants continued to show a significant preference for the options that frequently produced better outcomes, although the effect was weaker. Model comparison indicated that the RANGE-FREQUENCY model provided the best fit to the data, and we once again showed that the FREQUENCY and RANGE-FREQUENCY models were the only ones capable of reproducing the qualitative patterns of results using only the optimized parameters.

Taken together, the choice results across both experiments strongly support RL models that incorporate frequency encoding. According to these models, the subjective value of an outcome is partially determined by its rank

within a sample of recently experienced outcomes from the same context (Stewart et al., 2006). A consequence of frequency or rank encoding is a general preference for options associated with better local outcomes. It is a well-established finding in the experience-based choice literature that people are attracted to options that produce the best outcomes most of the time, even when those options have lower expected payoffs (e.g., Barron & Erev, 2003; Hayes & Wedell, 2021; Hertwig et al., 2004; Yechiam & Busemeyer, 2005). The most common explanation of this phenomenon is that people make decisions based on *small* samples of past experiences, which results in rare outcomes being underweighted relative to frequent outcomes (Erev et al., 2017; Erev & Barron, 2005). However, most versions of this theory still assume a context-independent subjective value function. This means that the small samples theory would not be able to explain why participants in Experiment 1 developed strong preferences for certain target options over others, as all three target options produced the same exact outcomes should thus be associated with the same subjective values. An advantage of the FREQUENCY and RANGE-FREQUENCY models is that they can account for underweighting of rare events and context-dependent preferences using a single mechanism, giving them greater generalizability.

In addition to demonstrating that the frequency encoding models outperformed the other context-dependent encoding models, we were able to rule out specific alternative theories in each experiment. First and foremost, our results clearly falsified the basic Q-learning model, which assumes absolute value representations. Participants in both experiments made choices that could

not be explained if they were encoding outcomes on an absolute scale, consistent with prior studies (Palminteri & Lebreton, 2021). Other theories have asserted that habits (i.e., stimulus-response associations) play a key role in guiding decisions from experience and that in certain situations, habits may exert greater control over choice than subjective values (i.e., stimulus-outcome associations) (Miller et al., 2019). While we would agree that habitual processes likely contribute to these types of repeated decisions (especially when they occur over longer timescales; Bavard et al., 2021), our results in Experiment 1 suggest that their influence in this case was minimal at best. Specifically, we showed that the difference in choice rates for two target options during the learning phase was not consistently associated with the preference for one option over the other in the transfer phase, and thus habits could not account for the full pattern of context-dependent preferences that we observed. In Experiment 2, we included models with separate learning rates for positive versus negative prediction errors (Niv et al., 2012) and for confirmatory versus disconfirmatory outcomes (Palminteri, Lefebvre, et al., 2017) to test whether our results could be explained by learning asymmetries in response to different types of feedback. Although the second model was only slightly worse than the winning RANGE-FREQUENCY model in terms of the relative comparison criteria, neither of the alternative models was able to generate the context-dependent choice behavior that we observed due to their reliance on absolute encoding. Thus, while learning asymmetries are certainly relevant in many RL tasks, they were not sufficient to explain our findings on their own.

Our model space included many plausible context-dependent encoding mechanisms, but other variants of some of the mechanisms exist. For example, a close relative of range adaptation is the divisive normalization mechanism, according to which the value of a single reward is normalized by dividing it by the *sum*, rather than the range, of other rewards in the local context (Louie & Glimcher, 2012). Divisive normalization is thought to be a canonical neural computation across multiple brain regions (Carandini & Heeger, 2012) and is commonly used to model context dependence in domains outside of RL, including preferential choice (Louie et al., 2013). To the best of our knowledge, there have been very few attempts to incorporate divisive normalization into an RL model (for examples, see Bavard & Palminteri, 2021; Louie, 2021). We devised two versions of a divisive normalization RL model, one in which the so-called semisaturation parameter in the denominator of the value normalization function was constrained to be 1.0, and another in which this parameter was freely estimated (see Appendix A). The two versions were tested in both experiments and compared to the other models based on BIC values. The results showed that the divisive normalization models were outperformed by the winning models in the first (FREQUENCY) and second experiment (RANGE-FREQUENCY) (Table A.1). For the transfer phase in Experiment 1, both models failed to generate a significant aggregate preference for the positive skew, high mean target option (PHM_{30}) over the negative skew, high mean target option (NHM_{30}), which the FREQUENCY model was able to capture due to the outcomes from PHM_{30} having higher local ranks (Figure A.1). In Experiment 2,

the divisive normalization models were unable to reproduce a significant preference for the options that frequently yielded better outcomes in the learning phase (Figure A.2), and their simulated transfer phase choice patterns were less aligned with the observed data than the patterns generated by the frequency encoding models (Figures A.3 and A.4). Thus, divisive normalization does not appear to be capable of accounting for the context-dependent choice behavior that we observed in this experiment.

There are many interesting questions that remain open for future research. One question is whether decision makers might engage different context-dependent encoding mechanisms depending on features of the choice task. There is already some evidence that outcome encoding in RL can adapt to task demands and expectations (Juechems et al., 2021), and that it might depend on what participants are attending to during the learning phase (Hayes & Wedell, 2022). Based on this, it is possible that the RANGE model performed poorly in the present study in part because the ranges were held constant across contexts, making the range of outcomes less salient to participants. In contrast, studies that have found support for range adaptation models have manipulated the range so that certain contexts have a wider range of outcomes than others (Bavard et al., 2018, 2021). Future studies should search for ways of manipulating range and skew simultaneously while still allowing for a dissociation between models to determine whether this is a critical factor. Another interesting question concerns blocked versus interleaved presentation formats. In a blocked format, all trials for a given context are presented together during the learning

phase, whereas in an interleaved format, the trials for different contexts are randomly intermixed as they were in the present experiments. A previous study has shown that range adaptation effects in RL are enhanced when contexts are presented in a blocked format (Bavard et al., 2021). However, in a preliminary experiment using the choice task from Experiment 2, we found that frequency encoding effects were significantly more pronounced in the interleaved condition than in the blocked condition (results not presented here). This may be because the interleaved condition places a greater demand on memory, and prior research suggests that rank encoding is enhanced when memory demands make it harder to remember absolute stimulus values (Pettibone & Wedell, 2007; Wedell et al., 2020). This explanation is purely speculative and requires additional investigation.

Finally, an important goal for future studies will be to test some of the auxiliary assumptions made by the FREQUENCY and RANGE-FREQUENCY models. Both models assume that the subjective value of an outcome is a weighted combination of its range and frequency values. The FREQUENCY model computes range values at the global level (i.e., using the global range), so that context-dependent preferences are entirely driven by the frequency principle. The RANGE-FREQUENCY model, in contrast, computes range values at the local level using a dynamic range adaptation process (Bavard et al., 2021). However, an alternative possibility is that there are no range computations, and that the subjective value of an outcome is fully determined by how it ranks within both local and global context (see Mullett & Tunney, 2013, for evidence of

hierarchical, rank-based value representations in the human brain). This could be accomplished by a model that computes two separate frequency values, one based on local context and another based on global context, or by a model that allows previously experienced outcomes from other contexts to have a nonzero probability of being recruited in the comparison sample. The degree of context dependence exhibited by such a model would depend on the relative weighting of local and global inputs. A second assumption of the FREQUENCY and RANGE-FREQUENCY models is that recent outcomes are given more weight than earlier outcomes in the computation of frequency values. While recency effects are well-established (Barron & Erev, 2003; Yechiam & Busemeyer, 2005), another finding is that people have better memory for the most extreme outcomes in a given context (Madan et al., 2014, 2021). It is very likely that extreme outcomes should be given greater weight than less extreme outcomes in the computation of frequency values, but this is something that our choice tasks were not designed to test. Future studies should consider additional ways of manipulating the outcome distributions within or between contexts to test different theories about the outcome retrieval process. Ultimately, determining what constitutes the effective context and how contextual features are subjectively evaluated will lead to a better understanding of how people make decisions from experience.

REFERENCES

Adaval, R., & Monroe, K. B. (2002). Automatic construction and use of contextual information for product and price evaluations. *Journal of Consumer Research*, *28*(4), 572–588. https://doi.org/10.1086/338212

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215–233. https://doi.org/10.1002/bdm.443

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06781-2

Bavard, S., & Palminteri, S. (2021, September 29 − October 1). *Contrasting range normalization and divisive normalization in human reinforcement learning* [Conference presentation]. Society for NeuroEconomics 19th Annual Meeting.

Bavard, S., Rustichini, A., & Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances*, *7*(14). https://doi.org/10.1126/sciadv.abe0340

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, *15*(1), 89–96. https://doi.org/10.3758/BF03205834

Burke, C. J., Baddeley, M., Tobler, P. N., & Schultz, W. (2016). Partial adaptation of obtained and observed value signals preserves information about gains and losses. *Journal of Neuroscience*, *36*(39), 10016–10025. https://doi.org/10.1523/JNEUROSCI.0487-16.2016

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. In *Nature Reviews Neuroscience* (Vol. 13, Issue 1, pp. 51–62). https://doi.org/10.1038/nrn3136

Choplin, J. M., & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General*, *131*(2), 270–286. https://doi.org/10.1037/0096-3445.131.2.270

Choplin, J. M., & Wedell, D. H. (2014). How many calories were in those hamburgers again? Distribution density biases recall of attribute values. *Judgment and Decision Making*, *9*(3), 243–258. http://journal.sjdm.org/13/13809/jdm13809.pdf

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. In *Current Opinion in Neurobiology* (Vol. 18, Issue 2, pp. 185–196). https://doi.org/10.1016/j.conb.2008.08.003

Don, H. J., & Worthy, D. A. (2021). Frequency effects in action versus value learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000896

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*(4), 912–931. https://doi.org/10.1037/0033-295X.112.4.912

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409. https://doi.org/10.1037/rev0000062

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cognitive, Affective and Behavioral Neuroscience*, *19*(3), 490–502. https://doi.org/10.3758/s13415-019-00723-1

Hayes, W. M., & Wedell, D. H. (2021). Regret in experience-based decisions: The effects of expected value differences and mixed gains and losses. *Decision*, *8*(4), 277–294. https://doi.org/10.1037/dec0000156

Hayes, W. M., & Wedell, D. H. (in press). Reinforcement learning in and out of context: The effects of attentional focus*. Journal of Experimental Psychology*: *Learning, Memory, and Cognition.*

Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*(1), 26–38. https://doi.org/10.1037/h0040013

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological*

Science, *15*, 534–539. http://library.mpib-berlin.mpg.de/ft/rh/RH_Decisions_2004.pdf

Higgins, E. T., & Lurie, L. (1983). Context, categorization, and recall: The "change-of-standard" effect. *Cognitive Psychology*, *15*(4), 525–547. https://doi.org/10.1016/0010-0285(83)90018-X

Hunter, L. E., & Daw, N. D. (2021). Context-sensitive valuation and learning. *Current Opinion in Behavioral Sciences, 41*, 122–127. https://doi.org/10.1016/j.cobeha.2021.05.001

Juechems, K., Altun, T., Hira, R., & Jarvstad, A. (2021). Human value learning and representation reflects rational adaptation to task demands. *PsyArXiv*. https://doi.org/10.31234/osf.io/4vdhw

Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms16033

Kobayashi, S., De Carvalho, O. P., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience*, *30*(2), 534–544. https://doi.org/10.1523/JNEUROSCI.4009-09.2010

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006973

Louie, K. (2021). Asymmetric and adaptive reward coding arises from normalized reinforcement learning. *BioRxiv.* https://doi.org/10.1101/2021.11.24.469880

Louie, K., & Glimcher, P. W. (2012). Efficient coding and the neural representation of value. *Annals of the New York Academy of Sciences*, *1251*(1), 13–32. https://doi.org/10.1111/j.1749-6632.2012.06496.x

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 6139–6144. https://doi.org/10.1073/pnas.1217854110

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin and Review*, *21*(3), 629–636. https://doi.org/10.3758/s13423-013-0542-9

Madan, C. R., Spetch, M. L., Machado, F. M. D. S., Mason, A., & Ludvig, E. A. (2021). Encoding context determines risky choice. *Psychological Science*, *32*(5), 743–754. https://doi.org/10.1177/0956797620977516

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Mullett, T. L., & Tunney, R. J. (2013). Value representations by rank order in a distributed network of varying context dependency. *Brain and Cognition*, *82*(1), 76–83. https://doi.org/10.1016/j.bandc.2013.02.010

Niedrich, R. W., Sharma, S., & Wedell, D. H. (2001). Reference price and price perceptions: A comparison of alternative models. *Journal of Consumer Research*, *28*(3), 339–354. https://doi.org/10.1086/323726

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, *29*(44), 14004–14014. https://doi.org/10.1523/JNEUROSCI.3751-09.2009

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9096

Palminteri, S., & Lebreton, M. (2021). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, *41*, 144–151. https://doi.org/10.1016/j.cobeha.2021.06.006

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418.

Parducci, A. (1968). The relativism of absolute judgements. *Scientific American*, *219*(6), 84–90. https://doi.org/10.1038/scientificamerican1268-84

Pettibone, J. C., & Wedell, D. H. (2007). Of gnomes and leprechauns: The recruitment of recent and categorical contexts in social judgment. *Acta Psychologica*, *125*(3), 361–389. https://doi.org/10.1016/j.actpsy.2006.10.004

Pischedda, D., Palminteri, S., & Coricelli, G. (2020). The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *Journal of Neuroscience*, *40*(16), 3268–3277. https://doi.org/10.1523/JNEUROSCI.1712-19.2020

Pompilio, L., & Kacelnik, A. (2010). Context-dependent utility overrides absolute memory as a determinant of choice. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 508–512. https://doi.org/10.1073/pnas.0907250107

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Rangel, A., & Clithero, J. A. (2012). Value normalization in decision making: Theory and evidence. In *Current Opinion in Neurobiology* (Vol. 22, Issue 6, pp. 970–981). Elsevier Ltd. https://doi.org/10.1016/j.conb.2012.07.011

Riskey, D. R., Parducci, A., & Beauchamp, G. K. (1979). Effects of context in judgments of sweetness and pleasantness. *Perception & Psychophysics*, *26*(3), 171–176. https://doi.org/10.3758/BF03199865

Seymour, B., & McClure, S. M. (2008). Anchors, scales and the relative coding of value in the brain. *Current Opinion in Neurobiology*, *18*(2), 173–178. https://doi.org/10.1016/j.conb.2008.07.010

Smith, R. H., Diener, E., & Wedell, D. H. (1989). Intrapersonal and social comparison determinants of happiness: A range-frequency analysis. *Journal*

of Personality and Social Psychology, *56*(3), 317–325. https://doi.org/10.1037/0022-3514.56.3.317

Soukupová, M., Garcia, B., & Palminteri, S. (2021, June 10 − 11). *Context-dependence induces false memories of economic values: A test across three modalities and four preference elicitation methods* [Conference presentation]. 11th Annual Interdisciplinary Symposium on Decision Neuroscience.

Steingroever, H., Wetzels, R., & Wagenmakers, E. J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *1*(3), 161–183. https://doi.org/10.1037/dec0000005

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881–911. https://doi.org/10.1037/0033-295X.112.4.881

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–1645. https://doi.org/10.1126/science.1105370

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*, 704–708. www.nature.com

Tripp, J., & Brown, G. D. A. (2016). Being paid relatively well most of the time: Negatively skewed payments are more satisfying. *Memory and Cognition*, *44*(6), 966–973. https://doi.org/10.3758/s13421-016-0604-0

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Roherer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 297−294). New York: Harper & Row.

Wedell, D. H. (1996). A constructive-associative model of the contextual dependence of unidimensional similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 634–661. https://doi.org/10.1037/0096-1523.22.3.634

Wedell, D. H., Hayes, W. M., & Kim, J. (2020). Context effects on reproduced magnitudes from short-term and long-term memory. *Attention, Perception, and Psychophysics*. https://doi.org/10.3758/s13414-019-01932-z

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology*, *55*(3), 341–356. https://doi.org/10.1037/0022-3514.55.3.341

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *The American Journal of Psychology*, *102*(2), 233. https://doi.org/10.2307/1422955

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin and Review*, *12*, 387–402. https://doi.org/10.3758/BF03193783

APPENDIX A

ADDITIONAL MODELS

Here, we describe a model that uses divisive normalization to encode outcomes on a context-dependent scale. Our formulation of the divisive normalization mechanism was based on Louie et al. (2013). The subjective value of the *i*th outcome on trial *t* is computed as follows:

$$v_{i,t} = \frac{r_{i,t}}{\sigma^2 + \omega \cdot \sum_{k=1}^{K} r_{k,t}}$$

where σ^2 is called the semisaturation parameter and ω is the normalization weight. The normalization term in the denominator is simply the sum of the *K* outcomes presented on trial *t*. When ω equals 0, the normalization term vanishes and subjective values are a linear function of objective outcomes. When ω equals 1, subjective values are a nonlinear function of objective outcomes. Because we were not sure what effect σ^2 would have in the present study, we tested two versions of the model, one in which σ^2 was fixed to 1.0 and another in which σ^2 was freely estimated [Louie et al. (2013) reported that σ^2 was not necessary to account for context dependence in their study]. Both versions of the divisive normalization model used the delta learning rule to update reward expectations (Equation 1) and the softmax function to compute choice probabilities (Equation 3).

The divisive normalization models were compared to the winning models in both experiments using BIC values (Table A.1). The constrained version with σ^2 fixed to 1.0 was preferred over the full version in both cases, but both were outperformed by the FREQUENCY and RANGE-FREQUENCY models. Figures A1 through A4 show that the divisive normalization models failed to capture key behavioral signatures of frequency encoding in both experiments.

Table A.1. Additional Model Comparison Results

Model	Parameters	Experiment 1	Experiment 2
FREQUENCY	4	**195.46**	298.20
RANGE-FREQUENCY	5	196.38	**297.22**
DIVISIVE NORM v1	4	199.29	311.26***
DIVISIVE NORM v2	5	203.96**	316.41***

Note. Mean Bayesian information criterion (BIC) values. The best model in each experiment is shown in bold. Significance tests reflect comparisons of each model to the best model using paired t-tests (df = 59 and 49 in Experiment 1 and 2, respectively). BIC = $-2 \times LL + k \times \ln(n)$, where LL is the maximized log-likelihood, k is the number of model parameters, and n is the number of observations. The semisaturation parameter σ^2 was fixed to 1.0 in the first version of the divisive normalization model (DIVISIVE NORM v1) and freely estimated in the second (DIVISIVE NORM v2).
**$p < .01$., ***$p < .001$

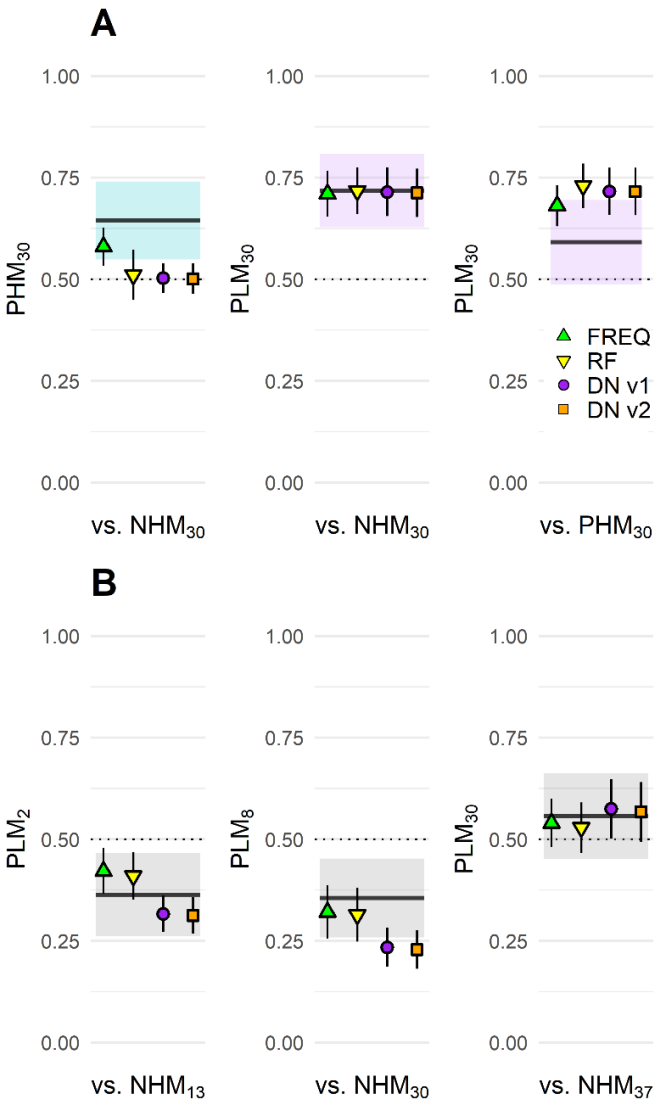


Figure A.1 Additional Model Simulations in Experiment 1. Pairwise choice preferences for the target pairs (A) and opposite skew pairs (B). Each panel shows the mean proportion of times the option on the vertical axis was chosen over the option on the horizontal axis, averaging across participants. The solid black lines and shaded boxes show the means and 95% confidence intervals for the observed data. The points and error bars show the means and 95% confidence intervals for the RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. FREQ = FREQUENCY. RF = RANGE-FREQUENCY. DN v1 = DIVISIVE NORM v1 (σ^2 = 1.0), DN v2 = DIVISIVE NORM v2.

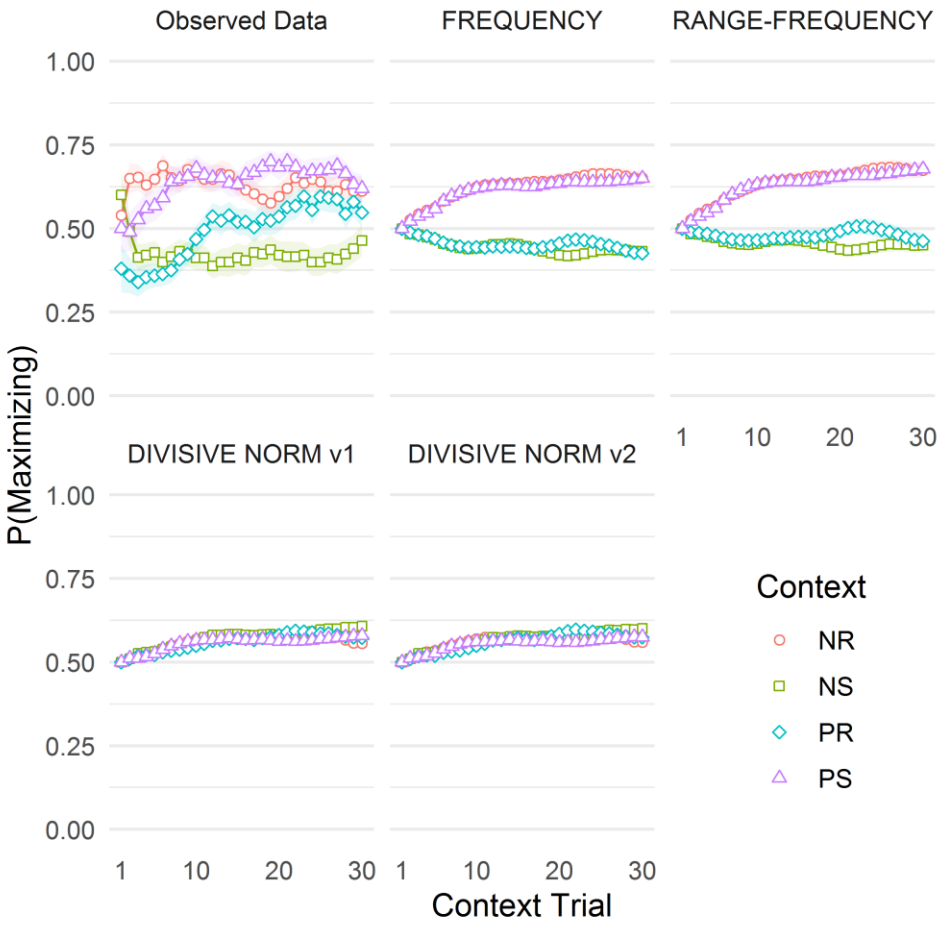


Figure A.2 Additional Model Simulations in Experiment 2: Learning Phase. Mean proportion of EV-maximizing choices across the 30 learning trials for each context. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. In all panels, choices were smoothed using a 5-trial rolling average prior to averaging across individuals. Error bands represent +/− 1 standard error.

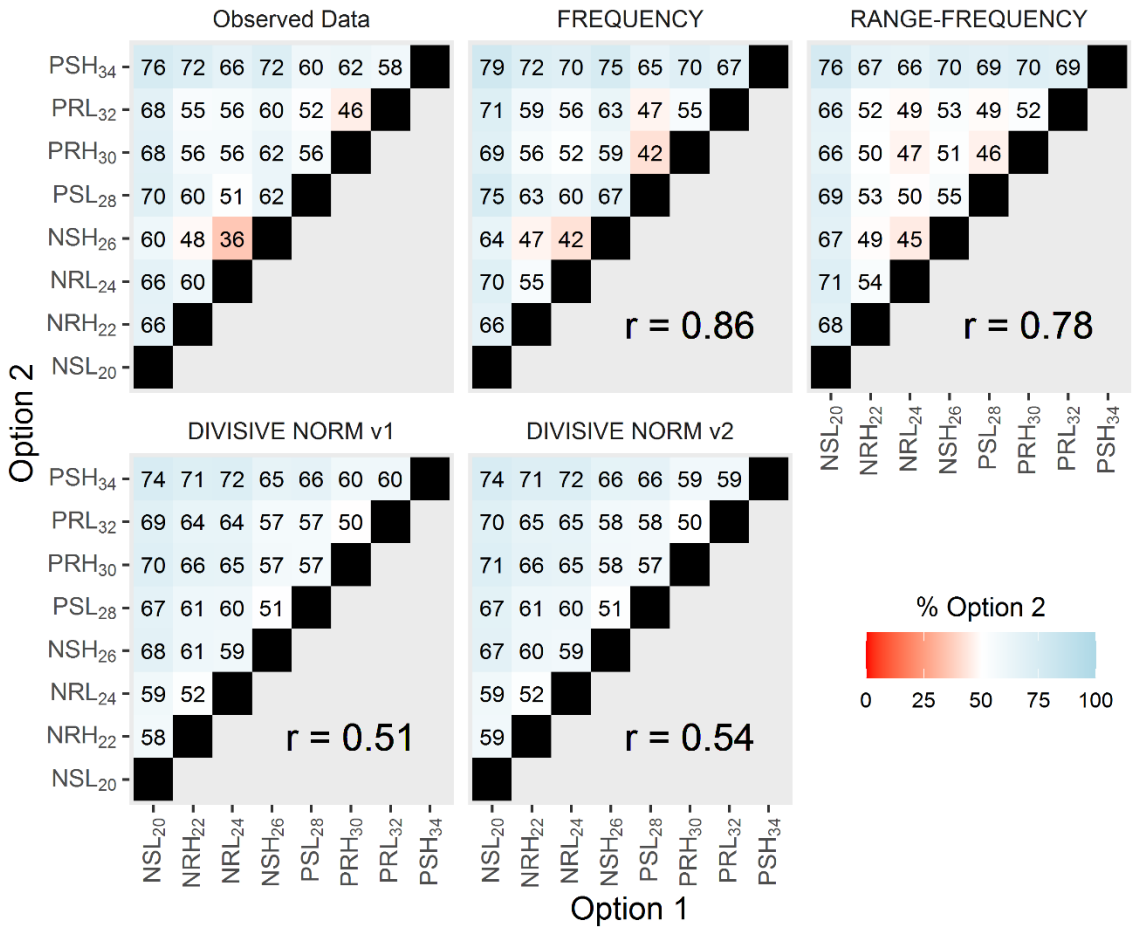


Figure A.3 Additional Model Simulations in Experiment 2: Transfer Patterns. Each cell shows the mean percentage of times the EV-maximizing option (Option 2) was selected, averaging across individuals. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. Also shown are the correlations between the empirical choice pattern and the patterns for each model.

Figure A.4 Additional Model Simulations in Experiment 2: Choice Rates. Mean choice rates for each option in the transfer phase, averaged across individuals. Choice rate is defined as the number of times an option was selected divided by the number of times it was presented. The top left panel shows the observed data, and the remaining panels show RL model simulations. Models were simulated using the fitted parameters for each participant and the results were averaged across 100 iterations. The models were not provided with participants' actual choices for the simulations. Error bars represent 95% confidence intervals.