Computational accounts for episodic memory in reinforcement learning

A Dissertation

Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University

Department of Psychology

Angela Gutchess, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

John Ksander

August 2020

The signed version of this form is on file in the Graduate School of Arts and Sciences.

This dissertation, directed and approved by John Ksander's Committee, has been accepted and approved by the Faculty of Brandeis University in partial fulfillment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Eric Chasalow, Dean
Graduate School of Arts and Sciences

Dissertation Committee:

Angela Gutchess, Department of Psychology
Paul DiZio, Department of Psychology
Donald Katz, Department of Psychology
Christopher Madan, Department of Psychology, University of Nottingham

ACKNOWLEDGEMENTS

I want first to thank my mom, Maura McGrane. Thank you for being a strong role model, and for teaching me the value of enjoying your work. Your support and encouragement has made all the difference for me.

I want to thank my advisor, Angela Gutchess, for the outstanding mentorship that made this possible. I sincerely appreciate your patience, guidance, and for showing me how to turn my interests into realistic projects. Thank you for making grad school a great experience!

I want to thank Paul Miller for an excellent four-year rotation. You have taught me a tremendous amount and given me new perspectives on science, problem-solving, and many other topics. Thank you for the opportunity and for being a great teacher — you really went above and beyond for me.

I also want to thank Chris Madan. Thank you for your enthusiasm, instruction, and tireless help over the years. Working with you before grad school made me excited about research, and excited about new and creative ideas. That was an influential experience, and thank you for continuing to be a great mentor since then.

Finally, I want to thank everyone who gave me opportunities, helped me, taught me, and supported me before and throughout grad school. I consider myself fortunate to say that includes too many people to list here, and I'm grateful for knowing all of you. Thank you!

ABSTRACT

Computational accounts for episodic memory in reinforcement learning

A dissertation presented to the Faculty of the
Graduate School of Arts and Sciences of Brandeis University
Waltham, Massachusetts

By John Ksander

Episodic memory involvement in reward learning has become a recent focus in decision-making research. Although it seems intuitive that we may remember rewarding experiences differently from episodes without reward, this topic has proven difficult to study empirically. The existing literature features few studies and their collective findings are inconsistent. These mixed results now require a cohesive explanation. The current work proposes that people better remember contextually valuable experiences, and that mechanism provides a coherent explanation for the empirical data. Modeling evaluated this account using data from three prior studies that shared a common experimental paradigm but produced different memory outcomes. In these prior studies, participants first learned reward values associated with words through two-alternative forced choice (2AFC) tests. Participants were then asked to recall as many words as they could remember. Crucially, differences in the 2AFC test procedure produced three qualitatively different results: high-low, U-shaped, and attenuated relationships between reward and memory. These three results were previously interpreted in terms of additive psychological processes (e.g., reward, salience, and boundary effects) or stimulus competition in 2AFC pairings. In contrast, the current work tests whether all three results are explained by a model that assumes a memory's strength depends on its value. Reward and value are distinct concepts in this context,

and a memory's value measures how much the mnemonic information changes future reward expectations. I modeled this account within a reinforcement learning framework and simulated participants' learning and memory behavior during these previous studies. The model successfully reproduced all three reward-memory relationships observed in these experiments, without requiring additional psychological processes. The modeling effort also gave falsifiable experimental predictions about how changing 2AFC outcomes should restore the attenuated memory differences. Together, these results show a coherent mechanism that explains how reward influences memory.

Contents

List of Tables

List of Figures

1. Introduction

It seems intuitive that our ability to remember specific learning experiences can benefit our decision-making. Understanding the mechanisms behind that intuitions would greatly improve formal models of learning and memory. However, the relationship between decision-making and episodic memory has proven difficult to study empirically. The existing literature features few studies on this topic and their collective findings are inconsistent. Different experimental approaches have given mixed results that now require a cohesive explanation. This dissertation hypothesizes that a memory's strength depends on the mnemonic information's value, and that mechanism can provide a coherent explanation for the empirical data. Modeling will compare this account to alternative explanations in data from three separate studies.

1.1 Model framework for learning and memory

Reinforcement learning has become a ubiquitous framework for studying human behavior. With roots in both behavioral psychology (Rescorla & Wagner, 1972) and artificial intelligence (Sutton, 1988), reinforcement learning accounts have become notable for their explanatory power in psychology (Niv, 2009) and impressive ability to emulate human-like intelligence (Silver et al., 2016). Broadly, this framework describes learning in terms of experiencing prediction errors for anticipated outcomes, and adjusting subsequent predictions to minimize future errors.

A great deal of this research has focused on two fundamental strategies in reinforcement learning (RL): model-free and model-based learning. In both strategies, the model agent (i.e.,, learner) attempts to learn what experiences and behaviors will lead to rewarded outcomes. The agent's experiences comprise the observed stimuli or environment at a given moment, referred to as state observations. In model-free (MF) learning, the agent stores running average value-

estimates for states and actions. At decision time, the agent simply chooses the available option with the highest value-estimate. This roughly corresponds to “habitual learning”, or operant conditioning. Model-based (MB) learning, in contrast, corresponds more closely to goal-directed behavior and incorporates planning. A MB agent learns a model for the environment via the transition probabilities between states, essentially producing a cognitive-map for the task. At decision time the MB agent evaluates possible sequences of behaviors and their estimated consequences. The agent then selects the action corresponding to the best long-run behavioral strategy.

Human behavior reflects both MF and MB learning, and many studies have investigated why we alternate between these strategies (Daw, Niv, & Dayan, 2005) and their possible neural mechanisms (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Glascher, Daw, Dayan, & O'Doherty, 2010; Wunderlich, Smittenaar, & Dolan, 2012). Adding to this, recent evidence shows our behavior also reflects successor representation learning (Momennejad et al., 2017); a distinct third algorithm for learning predictive sequences within a task. The successor representation (SR) learns long-run state trajectories through the environment by updating running average state expectations (Gershman, 2018). In this way, the SR can be thought of as incorporating aspects of both MF and MB algorithms. Successor representations learn to anticipate future experiences similarly to how MF agents learn values (i.e., through something like “habitual learning”). This produces a cognitive map for the task, but without using the transition probabilities between all pairs of states (i.e., as in MB learning). The SR lies somewhere in between MF and MB algorithms by predictive mapping through slow, aggregate learning.

None of these algorithms possess a human-like episodic memory system, although the SR or the MB learner's world-model could be considered a semantic memory for the environment. However, SR learning does provide the fundamental computations needed for episodic memory. Gershman, Moore, Todd, Norman, and Sederberg (2012) show how the successor representation computations are functionally equivalent to those in the Temporal Context Model (TCM; Howard & Kahana, 2002). The general idea behind this connection is that SR learns discounted future state-occupancies through recency-weighted state observations, and such learning encodes an experience's temporal context. That is, SR updates are based on simultaneously observing the current state, along with previous states weighted by their recency. These updates therefore encode an experience along with its temporal context, and context encoding forms the basis of episodic memory in TCM.

1.2 What use is an episodic memory?

Mastering games such as StarCraft (Vinyals et al., 2019) and Go (Silver et al., 2016) highlight the remarkably intelligent behavior possible with reinforcement learning. Also striking, the sophisticated modeling behind these accomplishments lack mechanisms resembling human episodic memory. Such behavior may be possible without episodic memory at all. In light of similar observations, Lengyel and Dayan (2008) poignantly asked: what use is an episodic memory? Answering this question may address some of the many remaining problems with how RL produces and explains intelligent behavior. As such, this topic has become a recent focus point in the decision-making literature (for an extensive review, see Gershman and Daw (2017)). Our behavior is heavily influenced by our memories of past events, so understanding memory within the RL framework (or vice versa) may provide key insights into learning, decision-making and memory.

Lengyel and Dayan (2008) provided a theoretical demonstration of how a basic episodic memory system can compensate for the shortcomings of MF and MB learning. In their simulations, the model learner with episodic memory (episodic-model) simply stored the previous sequence of actions and states when the agent experienced a larger-than-expected reward. Whenever the episodic-model agent reencountered a state from one of those "memories", the agent would attempt to follow that same sequence. Lengyel and Dayan (2008) observed the episodic-model outperformed both MB and MF when the learners had limited prior experience, or had difficulty learning a cognitive-map for the environment. These observations have informed much of the subsequent research on this topic (as reviewed in Gershman and Daw (2017)).

Decision-making in novel situations presents a difficult problem, and one where episodic memories may be particularly advantageous. In these scenarios, we have very little prior experience to draw from, and those experiences may be disjointed or unrepresentative of the larger task-space. There is not enough knowledge for careful goal-directed planning, nor experience for appropriate habitual behavior. However, simply remembering a successful episode and mimicking that memory can lead to reasonable decision-making in these circumstances (Lengyel & Dayan, 2008). Furthermore, natural task-spaces are seldom static, perfectly observable, or otherwise ideal for learning an accurate cognitive map. Even if humans had computer-like abilities to evaluate thousands of possible actions and their outcomes before making a decision, such planning completely relies on accurate task-space knowledge. Lengyel and Dayan (2008) shows that memory may be advantageous for situations where learners have some prior experience, but do not fully understand task-space yet.

To this end, guiding behavior with memory may also prove advantageous when one cannot easily observe the task-space. Poker provides a good example for how decision-making becomes more difficult in these circumstances. Decision-making in this game depends on whether players believe their own cards are higher-valued than their opponents' cards. This judgement is difficult because players cannot see their opponents' cards; the task is only partially observable. Fully observable problems are certainly the exception rather than the rule in naturalistic settings, and this poses a problem for RL approaches. The vast majority of RL research concerns fully observable Markov decision processes (MDP), and learning performance degrades significantly in partially observable tasks (Sutton & Barto, 2018). This highlights another shortcoming in how fundamental RL strategies (i.e., MB and MF) account for human behavior, and episodic memory may address this gap as well (Zilli & Hasselmo, 2008a, 2008b).

1.3 Measuring memory during decisions

The previously described lines of research have convincingly shown decision-making should involve memories for specific past experiences, and perhaps critically in certain circumstances. However, there is currently little direct evidence for such involvement (Gershman & Daw, 2017). Reinforcement learning studies typically use paradigms that are not well suited for discerning memory-driven outcomes. Subjects in these studies often learn reward outcomes from repeating many highly similar experiences. These designs are poorly situated for measuring unique memories about specific experiences. Furthermore, MB and MF learning describe pervasive human learning strategies, and account for a great deal of decision-making behavior. It is therefore difficult to show memory-driven decisions that are unambiguously not due to MB or MF learning.

Some studies have used artificial, but effective, experimental paradigms to isolate how episodic memories influence decisions. Barron, Dolan, and Behrens (2013) asked subjects to decide between two food options, each comprised of food already familiar to the subject but combined in a novel way. For instance, subjects decided whether they would prefer "tea-jelly" or "avocado-raspberry smoothie". The fMRI evidence in Barron et al. (2013) indicated subjects estimated the novel stimulus values (e.g., "tea-jelly") with their memories for the familiar component stimuli (e.g., tea, and jelly). In another study by Gluth, Sommer, Rieskamp, and Buchel (2015), subjects abstained from food and then rated their preference for a series of snack items. Afterwards, in an fMRI scanner subjects saw images of the snack items at different screen locations (e.g., snack images located in a 2 x 6 grid). Subjects were subsequently probed with two screen locations, and asked which snack they preferred. Crucially, in order to obtain the preferred snack, the subject must first remember both options using the location cues. Gluth et al. (2015) found subjects' decisions relied on how well they could remember the snacks, and that subjects were in fact biased to choose items with higher memory strength (when controlling for snack value).

In both the Barron et al. (2013) and Gluth et al. (2015) paradigms, the task construction forced subjects to make decisions based on their episodic memories. The decision-behavior directly reflects contributions from memory in these circumstances. However, most studies on memory and decision-making indirectly measure how memories relate to decisions. Prior studies have typically linked memory to decision-behavior using sperate memory tests outside the decision-making task (Madan & Spetch, 2012; Rouhani, Norman, & Niv, 2018; Wimmer, Braun, Daw, & Shohamy, 2014). A few experiments have manipulated memory directly, for example, by changing contextual familiarity (Duncan & Shohamy, 2016) or recall training tasks (St-

Amand, Sheldon, & Otto, 2018). In two studies, Bornstein, Khaw, Shohamy, and Daw (2017) and Bornstein and Norman (2017) presented memory cues during a choice task. Both studies found cueing a memory prior to a choice trial biased subjects' decisions towards the remembered values. These results provide strong evidence that episodic memories can influence decisions. However, the mnemonic influences in Bornstein et al. (2017) and Bornstein and Norman (2017) are still incidental; subjects' decisions do not require the cued memory, and the memory recall itself is unrelated to the decision (aside from temporal proximity).

1.4 Unclear how reward influences memory

The exact relationship between reward magnitude, RPE, learning, and memory is unclear in the literature. It is clear some meaningful relationship exists, however different experimental procedures appear to yield mixed results. In one study Madan, Fujiwara, Gerson, and Caplan (2012) found better memory for high reward items. However, subjects had more difficulty relearning those high-reward items for a different memory task afterwards. In a follow-up study, Madan and Spetch (2012) used the same decision-task with a finer reward scale and found better memory for both the highest and lowest valued items, indicating extreme reward values strengthen encoding (also see: Ludvig, Madan, McMillan, Xu, and Spetch (2018)). In a task with continuous reward values, Rouhani et al. (2018) found absolute RPE predicted better subsequent recognition memory performance, similarly indicating both high and low prediction errors improve encoding. Absolute RPE was also associated with higher learning rates. Interestingly, Rouhani et al. (2018) reported subjects' item memory and learning rates were uncorrelated, despite observing a positive correlation between RPE and these measures independently. Together, these studies indicate either extreme prediction errors or rewards improve memory

encoding and learning. However, independent prediction error and reward effects may exist, and rewarded information may be difficult to relearn for other purposes.

However, other findings show the relationship between reward and subsequent memory differently. For instance, using only two reward levels (i.e., high and low) Murty, FeldmanHall, Hunter, Phelps, and Davachi (2016) found item recognition alone did not correlate with better decision-making performance. Subjects were not more likely to choose a high-reward item over another low-reward item, even when they were more likely to recognize that high-reward item on a subsequent memory test. Jang, Nassar, Dillon, and Frank (2019) designed a tightly-controlled gambling task to disassociate reward from its associated experiences (e.g., surprise, uncertainty). They found only positively-signed RPE strengthened memory encoding, and that reward magnitude showed no relationship with memory. In a behavioral and fMRI study, Wimmer et al. (2014) reported memory performance was actually negatively correlated with learning rate, reward, and RPE. That is, Wimmer et al. (2014) found better learning, encountering large rewards, or experiencing large prediction errors all resulted in poorer memory. The authors concluded there may be a competition between error-based learning and memory formation in striatal functioning. All together, these results are difficult to collectively reconcile, even when considering the procedural differences across studies.

1.5 Reward versus value

The current proposal treats reward and value as distinct concepts, as they are defined in reinforcement leaning (Sutton & Barto, 2018). The distinction between reward and value is critical for this proposal. Prior studies have mostly focused on how reward influences subsequent memory. However, their results are difficult to reconcile from this perspective. I propose these studies are better understood through how *value* influences memory.

Reward is the learner's goal; money or points represent typical rewards in human psychology experiments. Value indicates an expectation of reward over a longer term, and this expectation incorporates elements such as environmental dynamics and the learner's behavior. Consider a hypothetical scenario in which you are going out to a restaurant for dinner. Food is the reward in this scenario. First you arrive at the restaurant, and then someone seats you at a table. At this point you have received zero reward, however your situation's value has increased because you are now closer to eating dinner. This proposal's hypothesis concerns a memory's value, or how information in memory relates to reward expectations.

1.6 Hypothesis

The existing literature shows a complicated relationship between decision-making and memory. Different experimental procedures have yielded mixed findings that now require a cohesive explanation. I hypothesize that a memory's strength depends on the mnemonic information's value, and that mechanism provides a unifying account for the experimental data. I will test this hypothesis by simulating previous studies with different models and comparing their behavioral predictions to experimental data. I predict that models assuming subjects better remember valuable information will best fit the empirical data. This will provide a coherent mechanism that explains reward-memory effects in the literature.

2. Reward-memory outcomes in simple situations

Madan, Fujiwara, Gerson, and Caplan (2012) examined how well people remembered stimuli after a value-learning task. Subjects first learned word values through a 2 alternate-forced-choice (AFC) test, where they chose words and received associated rewards. These subjects were then asked to recall as many words as they could remember, and completed an additional implicit memory task. The authors found subjects recalled high-reward items more frequently, and responded more quickly to high-reward items during the implicit memory task. In a second experiment, Madan et al. (2012) also found subjects experienced greater difficulty relearning high-reward items in a new context. These results indicated that rewards in the value-learning task influenced subsequent memory. Moreover, the data showed stronger memories for higher-reward items.

These experiments are notable for providing the first data showing reward influences memory with unrewarded memory tests. Many studies previously examined how people prioritize learning in preparation for rewarded memory tests (Adcock, Thangavel, Whitfield-Gabrieli, Knutson, & Gabrieli, 2006; Harley, 1965). Rewarded tests motivate subjects to study certain items by promising rewards for remembering those items in the future. Studies on motivated learning generally show how people can strategically encode memories that maximize reward at test (Castel, Benjamin, Craik, & Watkins, 2002). However, such data does not necessary inform how people remember rewarding experiences or value-learning episodes.

Madan et al. (2012) employed a simple value-learning task, particularly appropriate for an initial study on this topic. Each word stimulus was assigned either 1-point or 10-points (points later redeemed for money). Subjects saw two words simultaneously, chose one word and received the associated point reward. This represents a rare situation when reward and value are

essentially equivalent (see section 1.7). Subjects learn through a series of simple independent decisions, where the value of choosing a word equals the number of points immediately received. This certainly almost never occurs in naturalistic settings, but the simplicity provides an ideal starting point for this dissertation.

In the first experiment, subjects completed a lexical decision (LD) task following value-learning. In the LD task subjects indicated whether stimuli were valid English words. Words from the value-learning task, a novel word set, and non-word letter strings comprised the LD stimuli set. Faster reaction times (RT) for lexical decisions are considered to indicate stronger implicit memory (as reviewed by Schacter, Chiu, and Ochsner (1993)). Following the LD task, subjects completed a free recall task where they were simply asked to recall as many words from the study as possible.

The second experiment started with the same value-learning task. Afterwards, subjects performed a different memory task consisting of several study/recall cycles. In each cycle, subjects first studied a list containing both words from value-learning and new words; subjects were then asked to recall only the words from this list. This task is challenging because subjects have already encoded strong contexts for the value-learning words. They must associate these words with new a context, and then constrain their recall to that new context. Performance on this task was measured by both correct recall and intrusion error probabilities. Subjects commit an intrusion error when they recall a word from the experiment, but that word was not studied on the previous list. These errors measure how well subjects can constrain recall to only words from the previous list.

I hypothesized that more valuable experiences are easier to remember, and that mechanistically explains relevant data across multiple studies. Considered alone, the Madan et

al. (2012) memory results should be straight-forward because the simple learning task equates reward and value. High-reward words are always more valuable than low-reward words, so higher reward should always give stronger memories. I will simulate the experiments in Madan et al. (2012) with a model specifying that hypothesis. I will also simulate models that lack the hypothesized mechanism (i.e., models where memory does not explicitly depend on value). The results will most strongly support the hypothesis if the former model succeeds in reproducing the results, but the latter models fail. This would show the model can only account for the experimental data when memory depends on value.

2.1 Methods

2.1.1 Basic modeling description

I simulated the Madan et al. (2012) experiments using a reinforcement learning model (RL) with successor representation (SR; see section 1.1 for overview). I used a SARSA approach for temporal difference (TD) learning (Sutton & Barto, 2018). Broadly, SARSA methods model learning in terms of state-action-reward sequences (i.e., SARSA is "state-action-reward-state-action"), and learn a value function for state-action pairs. A state-action pair represents observing the current environment and taking an action. For example, a state-action pair in the Madan et al. (2012) value learning procedure might be "I see the words duck and house", followed by "I choose duck".

The notation in this section closely follows the notation in Sutton and Barto (2018). As such, uppercase letters indicate random variables and the corresponding lowercase letters indicate specific instances. For example, A_t and R_t note the action and reward on timestep t, and the lowercase a and r indicate a specific possible action and reward. This avoids excessive subscript indexing. Vectors are always $d\ x\ 1$ column vectors, and superscript $^{\mathrm{T}}$ notes transpose. I

will also highlight some differences from Sutton and Barto (2018) notation for clarity. For instance, I do not distinguish matrices or feature-representations with bolding. I also do not write the feature-representation x exclusively in lowercase (i.e., X indicates a random variable).

The following describes a general model update for timestep *t*.

$$E \rightarrow \lambda E + X_t \tag{1}$$

$$U \rightarrow U + \alpha_U X_t (R_t - X_t^T U) \tag{2}$$

$$W \rightarrow W + \alpha_W X_t [W^T (\gamma X_{t+1} - X_t) + E]^T \tag{3}$$

Equations 1-3 show updates for the eligibility trace *E,* reward expectation *U*, and successor representation *W*. For simplicity I am noting the state-action representation as X_t , rather than typical (S_t , A_t) notation.

This model represents states and actions with binary features. More specifically, the vector X contains an element for every feature in the simulation. The set of features includes both features for state observations, and actions. The state features indicate whether a stimulus is observed, and action features indicate whether a stimulus is chosen. If there are *n* total word stimuli, then $X(s, a) = \{s_1 \dots s_n ,\ a_1 \dots a_n\}$ contains a feature for observing every word and choosing every word. The state and action features for the i^{th} word are

$$s_i = \begin{cases} 1, & observed \\ 0, & otherwise \end{cases}$$

$$a_i = \begin{cases} 1, & chosen \\ 0, & otherwise \end{cases}$$

In this modeling approach (i.e., linear function approximation *(Sutton & Barto, 2018)),* the model cannot learn value information about feature interactions without those interaction terms explicitly included. For example, the model cannot learn how simultaneously observing words i and j changes the value function, without a feature representing both stimuli. I also included

such features because the value-learning task involves comparing two stimuli simultaneously. The set of features also included

$$s_{ij} = \begin{cases} 1, & observed\ both\ i\ and\ j \\ 0, & otherwise \end{cases}$$

In total, $X(s,a) = \left\{1\ ,\ s_1 \ldots s_n\ ,\ s_{12} \ldots s_{\binom{n}{k=2}}\ ,\ a_1 \ldots a_n\right\}$. X contains an intercept feature that always takes the value 1, a feature for observing each word, observing each combination of two words, and choosing each word. For clarity, the features relevant to each stimulus will be indexed as: observing the i[th] stimulus $x(s_i)$, simultaneously observing the i[th] and j[th] stimulus $x(s_{ij})$, and choosing the i[th] stimulus $x(a_i)$. In all feature vectors the intercept feature has the value 1 unless otherwise specified. The notation $x(s_i, a_i)$ indicates a feature vector for observing and choosing the i^{th} stimulus; $x(s_i, s_j, s_{ij}, a_i)$ indicates observing stimulus i and j while choosing stimulus i, etc.

Equation 1 shows the eligibility trace E updating to include the current experience X_t . An eligibility trace maintains decaying memories for previous experiences within the model, with decay specified by the γ parameter. Eligibility traces are therefore naturally analogous to psychological processes such as short term memory, and aspects of working memory (Todd, Niv, & Cohen, 2009).

Equations 2 and 3 show the updates for reward expectation *U* and successor representation *W.* Both terms update through TD error, and the difference between the respective error terms highlights their functional distinction. The reward expectation updates with $(R_t - X_t^T U)$, where R_t is the actual reward received for timepoint *t,* and $X_t^T U$ gives the predicted (scalar) reward; this error updates relevant features scaled by a learning rate $\alpha_U X_t (R_t - X_t^T U)$. In this way, U learns reward associations for features in X.

The successor representation updates with error term $(\gamma X_{t+1} - X_t)$, the discounted future features for state $t + 1$. The SR estimates future state trajectories, so the SR predicts a feature vector ($W^T X$) in this model because states have feature presentations (unlike predicting a scalar reward). Here the SR is an *F* by *F* matrix, where *F* is the number of total features. The W_{jk} entry gives the predicted (discounted) future encounters with the j^{th} feature, given an encounter with the k^{th} feature. In other words, *W* learns what features typically follow a given feature.

Value estimations q are given by:

$$q(x) = x^T WU \tag{4}$$

Here the SR distinguishes value estimations from reward expectations (see section 1.7). The underlying idea is that W predicts discounted future states, and the model learns W under a reward-maximizing behavioral policy. That is, the SR estimates the future states leading to maximum reward. Consider again the hypothetical scenario about dining out at restaurant. Food is the reward in this scenario, and you have just arrived at the restaurant. Many future states are possible (e.g., taking a phone call, using the restroom), but your successor representation should predict someone will seat you at a table. You learned this prediction because reward-maximizing restaurant behavior leads to a dining table more often than other possibilities. Again, a dining table is a highly valuable state despite totally lacking food (i.e., table has no reward expectation). The table's value is based on the prediction that food will quickly follow, and you have a reward expectation for food.

The current study models the Madan et al. (2012) experiments by simulating individual subjects' behavior on a trial-by-trial basis. These simulations included 99 subjects for the first experiment, and 72 subjects for the second experiment. The relevant summary statistics were

then calculated as described in Madan et al. (2012), and compared to the reported experimental statistics.

Table 1. Variables and descriptions.

variables	description	dimensions
F	Total number of features	scalar
E	Eligibility trace	F x 1
U	Reward expectation	F x 1
W	Successor representation	F x F
X	Feature representation	F x 1
X_0	Memory cue representation	F x 1
M	Word memory strengths	F x 1
λ	Decay parameter for Eligibility trace	Scalar
α_U	Learning rate for reward expectation	Scalar
α_W	Learning rate for SR	Scalar
γ	Discount factor	Scalar
σ_D	Drift rate variance (recall noise)	Scalar
θ_D	Evidence threshold	Scalar
Q	Value estimate(s)	-
a	Action (NOGO action with ∅ accent)	-
s	State	-

2.1.2 Value learning task

Individual simulated subjects[1] were initialized with all W, U , and E values set to zero before starting the value-learning task. Subjects completed 13 blocks of 18 trials each following the trial structure described in Madan et al. (2012). In each trial, subjects saw two words from the total stimulus set (36 words total) and chose one word. The model implements this behavior as choosing between two state-action representations that differ only in their action features (i.e., choosing an action). The model estimates the state-action values $Q_t(x)$ according to equation 4. A softmax behavioral policy then gives probabilistic decisions based on those value estimates.

A softmax policy gives the probability of taking action a given the currently observed state s as:

$$p(a|s) = \frac{\exp(q(s,a)/\beta^{-1})}{\sum_a \exp(q(s,a)/\beta^{-1})}$$

The connection to previous equations may be more clearly written as:

$$p(a|s) \;=\; p(X_t = x) \;=\; \frac{\exp(q(x)/\beta^{-1})}{\sum_{x \in X_t} \exp(q(x)/\beta^{-1})} \tag{5}$$

The softmax parameter β governs how strongly value estimates influence choice probabilities. Setting β close to zero will give uniform action probabilities regardless of value estimates ($p(x) \sim 1/n$). On the other hand, $\beta > 1$ will assign an action probability near 1 to the highest valued option, and probabilities near zero to the alternatives. I set $\beta = .95$ for all simulations in this dissertation.

Subjects received deterministic reward following each decision. The reward was either 1 or 10 points depending on whether the subject chose a low or high value word, respectively.

[1] "Subject" will refer to a simulated subject in the methods section, unless otherwise noted

Stimulus presentation followed the protocol described in Madan et al. (2012). Word pairs always contained one low and one high reward word, individual word pairings were randomized within each block, and each word was only shown once per block. Below shows the pseudocode for this task:

```
# prime notation distinguishes timestep variables ( X = X_t  ,  X' = X_{t+1} )
Initialize E, U, and W at zero
For each trial t … T:
    Q(X) = X^T W U    # estimate value for possible state-action pairs
    p(X = x) = exp(q(x)/β^{-1}) / Σ_{x∈X} exp(q(x)/β^{-1})    #softmax probabilities
    Choose x , observe r
    E ← λE + x
    U ← U + α_U x(r − xU)
    If t < T:
        W ← W + α_W x[W(γx' − x) + E]^T
    Else if t = T:
        W ← W + α_W x' E^T
```

2.1.3 Lexical decision task

A lexical decision (LD) task follows value-learning in Madan et al. (2012) experiment 1 (second experiment does not include LD). In lexical decision, subjects see 18 new words, 36 old words from value-learning, and 54 non-word letter strings. In each trial, subjects are shown a stimulus and must indicate whether it is a word or non-word letter string. Subjects complete eight practice trials before starting the task (included in the simulation).

Visual word recognition has a rich theoretical literature. Modern lexical decision models provide sophisticated accounts of reading behavior and successfully describe numerous related phenomena (as reviewed in Norris (2013)). However, those models are outside the current scope. The authors in Madan et al. (2012) used an LD task for measuring implicit memory, not as a means for studying reading behavior. Furthermore, the current modeling effort focuses on

episodic memory rather than implicit memory. This dissertation's main hypothesis is that a memory's strength depends on the mnemonic information's value. As stated, that hypothesis does not preclude implicit memories, but much less empirical data exists on that topic. The episodic memory data provides a far more rigorous test. Therefore, I simulated LD task behavior with a crude model that only reproduces the basic hypothesis; subjects show faster RT for more valuable items.

In each trial, stimulus values first are calculated according to equation 4 with stimulus representations $x(s_i, a_i)$ with intercept-feature values set to zero:

$$Q_i = x(s_i, a_i)^T WU \tag{6}$$

This means both state and action features contribute to stimulus values. In other words, stimulus values are the cumulative value of both experiencing and choosing the stimulus. The stimulus values were then simply rescaled into reaction times:

$$RT_i = \left(\frac{Q_i - \max(Q)}{\min(Q) - \max(Q)}\right)(a - b) + b \tag{7}$$

Where a and b are rescaling values meant to capture the longest and shortest RTs, respectively. These values were simply chosen to provide a realistic range for lexical decision reaction times, and were set to $a = 1000ms$ and $b = 400ms$ in all simulations. The presented stimulus' reaction time was then recorded from RT calculated in equation 7.

The updates in equations 1-3 were performed on each trial so that a subject's model reflected the LD task experiences. Notably, the model takes no actions during the LD task simulations, and this task was unrewarded. These updates served to simulate a subject's experience during the lexical decision task, rather than learn reward-maximizing behavior (the primary modeling objective in RL). Below shows the pseudocode for this task:

umlauts distinguish RT scaling variables in equation 7 from actions, etc

$r = 0$ # unrewarded task

For each trial $t \ldots T$:

$Q_i = x(s_i, a_i)^T WU$

$RT_i = \left(\frac{Q_i - \max(Q)}{\min(Q) - \max(Q)}\right)(\ddot{a} - \ddot{b}) + \ddot{b}$

here x is shorthand for $x(s_i)$ probed stimulus representation

$E \leftarrow \lambda E + x$

$U \leftarrow U + \alpha_U x(r - xU)$

If $t < T$:

$W \leftarrow W + \alpha_W x[W(\gamma x' - x) + E]^T$

Else if $t = T$:

$W \leftarrow W + \alpha_W x E^T$

2.1.4 Free recall

The exact memory task differed between experiments 1 and 2, although both experiments involved free recall. Free recall sessions in both experiments shared the same methodology. In response to a cue representation X_0 with the intercept-feature set to zero, the memory evidence for each stimulus was:

$$M_i = X_0^T W^T x(s_i, a_i) \tag{8}$$

In this equation $X_0^T W^T$ gives the memory evidence for all features in response to cue X_0. Just as in the value calculation (4), the memory evidence for each feature relevant to a stimulus comprises that stimulus' total evidence.

The free recall methods in this section closely follow the methodology Gershman et al. (2012) used to demonstrate successor representation equivalencies with TCM (see section 1.1). Gershman et al. (2012) simulated a passive study task followed by free recall, where $X_0^T W^T$ gave the memory strength for each stimulus. Their task did not include any actions or value estimations. In the current study, a stimulus memory comprises all relevant state and action features, extending the calculation to $X_0^T W^T x(s_i, a_i)$.

Memory recall then follows a linear ballistic accumulation (LBA; Brown and Heathcote (2008)) process with these evidence values. In LBA recall the memory evidence vector M was normalized to unit length and scaled by the coefficient of variation, following the recommendation in Gershman et al. (2012). The LBA recall process followed:

$$d_i \sim N(\mu = M_i\,,\ \ \sigma = \sigma_D)$$

$$b \sim U(0\,, A)$$

$$RT = t_0 + \frac{\theta_D - b}{d}$$

To reduce the number of free parameters during model fitting, I set $A = 0$ and $t_0 = 0$ for all simulations. This simplified the LBA recall to:

$$d_i \sim N(\mu = M_i\,,\ \ \sigma = \sigma_D)$$

$$RT = \frac{\theta_D}{d}$$

where d are stimulus drift rates, σ_D specifies recall noise, and θ_D specifies the evidence threshold. The recalled stimulus was simply the first to reach threshold: $\underset{i}{arg\,min}(RT_i)$. The recalled stimulus then provides the cue X_0 for the next recall. Additionally, the model always updates itself (1-3) following a successful recall, reflecting the subject's experience of "mental time travel" (Tulving, 1985).

2.1.5 Randomness in the model

In this model only action selection and free recall have stochastic terms, otherwise the model processes are deterministic[2]. Actions are chosen probabilistically via softmax, as discussed in section 2.1.2. However, the simulations in this dissertation always set β at $.95$ so that the action probabilities are always much greater for the highest-valued action. This means

[2] It is interesting to note that learning and memory encoding are noiseless in the model.

the model almost always chooses the highest-valued action, so noisy action-selections should not meaningfully influence the simulation results (i.e., not a meaningful source of variance).

On the other hand, free parameters (σ_D and θ_D) determine noisiness during free recall and recall noise contributes significantly to simulation outcomes. As described in section 2.1.4, σ_D specifies noise in evidence accumulation and θ_D specifies the evidence threshold. For clarity, changing the threshold determines how quickly memories are recalled, therefore also determining how many memories are recalled within the time limit. Together, the σ_D and θ_D parameters determine most variance in the simulation outcomes.

2.1.6 Simulations for Madan et al. (2012) Experiment 1 memory task

Subjects in the first experiment simply completed a five-minute free recall following the LD task. The pseudocode for this task was:

$X_0 = E$ # initialize the memory cue as the subject's current context (elig. trace)
$r = 0$ # unrewarded task
$t = 0$ # session time
$T = 300$ # five minute limit
$t_{delay} = .5$ # half second delay after response
While $t < T$:
    $M_i = X_0^T W^T x(s_i, a_i)$
    $d_i \sim N(M_i,\ \sigma_D)$
    $RT = \frac{\theta_D}{d}$
    $m = \arg\min_i (RT_i)$ # recalled stimulus index
    $t \leftarrow t + RT_m + t_{delay}$
    $x_m = x(s_m)$
    $E \leftarrow \lambda E + X_0$
    $U \leftarrow U + \alpha_U X_0 (r - X_0 U)$
    $W \leftarrow W + \alpha_W X_0 [W(\gamma x_m - X_0) + E]^T$
    $X_0 \leftarrow x_m$

2.1.7 Simulations for Madan et al. (2012) Experiment 2 memory task

Subjects in the second experiment completed six study-recall cycles. In each cycle, the subject studied 9 words comprised of: 3 one-point words, 3 ten-point words, and 3 new words. Subjects then attempted to recall only those words (i.e., words from the previous list) in a one-minute free recall. This study-recall cycle repeats 6 six times.

This memory task challenges human subjects because they must relearn words from value-learning in a new context, then constrain their recall specifically to that context. Furthermore, subjects strongly encoded value-learning words with their previous context (the value-learning task) over thirteen block repetitions. These aspects are similarly difficult for modeling subject behavior in the current study.

The learning rate parameters α_U and α_W largely determine subjects' performance on the value-learning task. High learning rates produce dramatic model updates following each experience. Large updates are useful in certain situations, particularly for learning quickly in stable, deterministic environments. Prediction errors in these situations are highly meaningful, and individual experiences accurately reflect the underlying dynamics. Conversely, large updates are maladaptive for learning in highly dynamic environments, or from stochastic outcomes. Individual prediction errors are less informative in volatile situations because specific experiences often poorly reflect the underlying dynamics. Lower α values enable learning in noisier situations by limiting how much a single event can influence the model, at the expense of slower learning (Preuschoff & Bossaerts, 2007).

When simulating the value-learning task with fixed value α_U and α_W parameters (i.e., non-changing parameters), only a certain range of α_U and α_W values produce successful learning behavior as seen in human subjects. These values must be high enough to enable learning within 13 blocks, but low enough to prevent erratic model updates that hinder learning. However, in the

subsequent memory task, a high α_W value is necessary to re-learn words in a new study list context. In particular, high α_W is necessary for learning a word's new association within one study presentation (i.e., necessary for quick learning). Preliminary modeling indicated that α parameters viable for successful value-learning behavior were not sufficiently high enough to subsequently re-learn contexts for the memory task. That is, fixed α_U and α_W values could not successfully reproduce both value-learning and memory task behaviors.

Evidence shows humans adjust "learning rates" as task characteristics change (Diederen & Schultz, 2015; Diederen, Spencer, Vestergaard, Fletcher, & Schultz, 2016), and it is entirely reasonable to assume human subjects in Madan et al. (2012) adapted their learning strategies during the experiment. One could easily justify selecting new α_U and α_W values prior to simulating the memory task. However, I addressed this issue with an adaptive learning rate method (Dabney & Barto, 2012) that provides a more satisfying explanation for how agents might adjust their learning rates. This method also uses fewer free parameters since the task specifies α rather than the experimenter.

The Dabney and Barto (2012) adaptive step-size method is also appealing for its intuitiveness. The underlying idea is that optimal α gives updates large enough to quickly minimize prediction errors (i.e., learn quickly), without overshooting the target. The implementation sets α for update t to the largest value that would not overestimate the target y_t with two consecutive updates. That is, repeating update t again on $t+1$ cannot give $\hat{y}_{t+1} > y_t$. I implemented this method to adjust α_W during experiment 2 simulations, while α_U remained fixed throughout the whole simulation (as in experiment 1 simulations). The implementation followed:

$$\alpha_t^W = \min[\alpha_{t-1}^W, |X_t^T(\gamma X_{t+1} - X_t)|^{-1}] \quad (9)$$

where α_W was initialized at 1 before each task in the simulation. Equation 9 corresponds to equation 15 in Dabney and Barto (2012). The prediction error term $X_t^T(\gamma X_{t+1} - X_t)$ differs here because equation 9 adapts the step-size for SR learning, where the prediction targets are state-action representations. Dabney and Barto (2012) illustrate their method in more typical SARSA learning with function approximation. All together, the pseudocode for the second experiment's memory task was:

$\alpha_0^W = 1$ #initialize SR learning rate at zero
$j = 0$ #index for adaptive step-size α_j^W
$E = 0$ # clear elig. trace to emulate distractor task
$r = 0$ # unrewarded task
$N = 9$ # words in each study list
$T = 60$ # one minute recall time limit
$t_{delay} = .5$ # half second delay after response
For each cycle $1 \ldots 6$:
    # list study
    For each word $n \ldots N$:
        $j \leftarrow j + 1$
        $E \leftarrow \lambda E + X$
        $U \leftarrow U + \alpha_U X(r - XU)$
        If $n < N$:
            $\alpha_j^W = \min\left[\alpha_{j-1}^W, |X^T(\gamma X' - X)|^{-1}\right]$
            $W \leftarrow W + \alpha_j^W X[W(\gamma X' - X) + E]^T$
        Else if $n = N$:
            $\alpha_j^W \leftarrow \alpha_{j-1}^W$
            $W \leftarrow W + \alpha_j^W x' E^T$
    # list recall
    $t = 0$
    $X_0 = E$ # initialize the memory cue as the subject's current context (elig. trace)
    While $t < T$:
        $M_i = X_0^T W^T x(s_i, a_i)$
        $d_i \sim N(M_i, \sigma_D)$
        $RT = \frac{\theta_D}{d}$
        $m = \arg\min_i (RT_i)$ # recalled stimulus index

$t \leftarrow t + RT_m + t_{delay}$

here x_m is shorthand for $x(s_m)$ recalled word representation

$j \leftarrow j + 1$

$\alpha_j^W = \min\left[\alpha_{j-1}^W, |X_0^T(\gamma x_m - X_0)|^{-1}\right]$

$E \leftarrow \lambda E + X_0$

$U \leftarrow U + \alpha_U X_0 (r - X_0 U)$

$W \leftarrow W + \alpha_j^W X_0 [W(\gamma x_m - X_0) + E]^T$

$X_0 \leftarrow x_m$

2.1.8 Parameter optimization

Model parameters were optimized with the Matlab particle swarm optimizer *particleswarm(),* using 200 particles for all searches. In-house modifications allowed the optimizer to interface with a SLURM scheduler and run particle simulations in a cluster computing environment. The particle swarm optimized $a_U,\ \gamma,\ \lambda,\ \sigma_D$, and θ_D for both experiments, in addition to a_W for experiment 1 (the second experiment set a_W via adaptive step-size algorithm). The objective score for experiment 1 was:

$$\epsilon_{learn}{}^3 + \epsilon_{LD}{}^3 + \epsilon_{recall}{}^3$$

where ϵ_{learn} , ϵ_{LD} , and ϵ_{recall} are the differences in summary statistics (i.e., average subject performance) between the simulated and empirical data for value-learning, lexical decision, and free recall. Due to the crude lexical decision modeling, the LD summary statistics were rescaled to zero-mean within calculating the error term. The purpose was capturing the pattern of results rather than replicating the precise empirical values, given the LD model's simple implementation. The objective score for experiment 2 was:

$$\epsilon_{learn}{}^3 + \epsilon_{recall}{}^3 + \epsilon_{int}{}^3$$

where ϵ_{int} difference in average subject intrusions between the simulated and empirical data. Summary statistics for the all relevant stimulus categories were included in error term

calculations. For instance, ϵ_{recall} sums the difference between simulated and empirical recalls for 10-point words, 1-point words, and new words. Individual error terms were exponentiated to favor solutions that work for all tasks.

2.1.9 Simulations with H_0 models

Additionally, I ran the same parameter search simulations with a model that lacks the terms explicitly specifying my hypothesis. That is, I reran the same simulations with a model that does not directly tie memory strength to value. These simulations test whether the model requires my hypothesis in order to reproduce the data in Madan et al. (2012). Specifically, in these simulations equation 6 was replaced with:

$$Q_i = x(s_i)^T WU \tag{10}$$

In this way, only state value dictates lexical decision RTs instead of state-action values. The value-learning task produces approximately uniform item values, since the value function updates both items (on a given trial) according to the trial's reward outcome. Said another way, receiving a 10-point reward updates both features for words seen during the trial; both the 10-point and 1-point words. Only action-features values differ meaningfully between words, intuitively because choosing words yields reward not simply observing them. Therefore, tying LD reaction times to item values alone actually removes value function's influence on RTs for 10-point words relative to 1-point words. The value function would still influence RTs for old words (seen during value-learning task) relative to new words, producing basic priming effects unrelated to the current study's hypothesis.

Equation 8 was also replaced with:

$$M_i = X_0^T W^T x(s_i) \tag{11}$$

such that item memory strengths were only tied to item successor representations, as in Gershman et al. (2012). This similarly removes the influence of values learned during the first task on memory strength. Aside from these two terms, the null simulations (and parameter searches) exactly matched the other simulations.

2.2 Results

2.2.1 Simulations for Madan et al. (2012) Experiment 1

Subjects were simulated with the best parameter values found during parameter optimization (Table 2). Figure 1 shows choice performance during the value learning task. Choice performance begins at chance and improves quickly in both the empirical and simulated data. In both datasets, the average subject chooses correctly on > 90% of trials by the final block. The qualitative similarity between learning curves in figure 1 indicates the model captures human subjects' choice performance extremely well in this task.

Figure 2 shows lexicial decision task RTs for 10-point (high), 1-point (low), and new words. In the empirical data, the key pattern was subjects responded fastest to high-reward words, slower to low-reward words, and the slowest to new (unrewarded) words; Madan et al. (2012) reports these differences were all statistically significant. The simulated data captures this pattern, but with different RT magnitudes (R^2 = -66.728, RMSE = 458.38). As previously mentioned in the methods section, LD task performance was crudely modelled, and the parameter search objective function specified the results pattern rather than exact magnitudes. However, the simulation succeeded in qualitatively matching the empirical data. In both experimental and simulated datasets, subjects respond fastest to high-reward words, slower to low-reward words, and the slowest to new words.

Table 2. Best parameters found for Madan et al. (2012).

Simulation	α_U	α_W	γ	λ	σ_D	θ_D
Experiment 1	$\sim N(.45, .04)$	.15	.971	.076	.125	3.597
Experiment 2	.346	-	.506	.633	.01	2.122
Experiment 1, H_0	$\sim N(.258, .04)$	.077	.73	.145	.015	3.991
Experiment 2, H_0	.272	-	.388	.571	.032	2.035

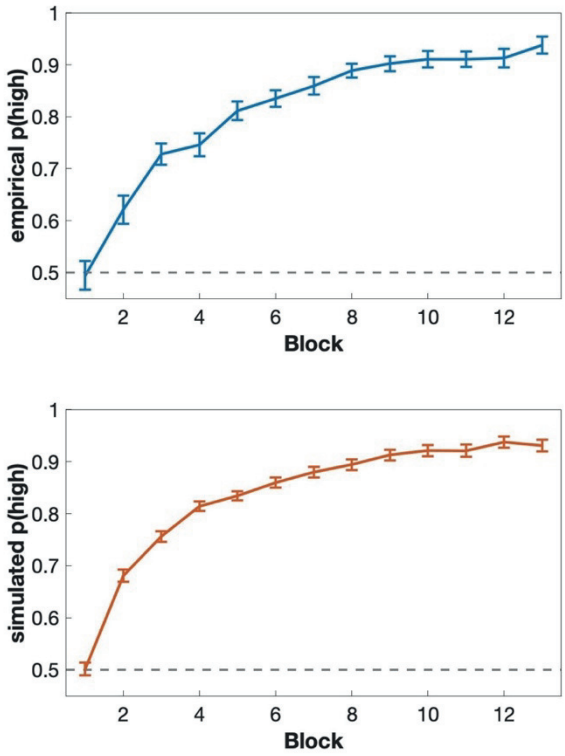

Figure 1. Value-learning in Madan et al. (2012) first experiment. p(high) shows likelihood subjects choose high-reward words over low-reward words. Top panel shows experimental data, bottom panel shows simulation data.

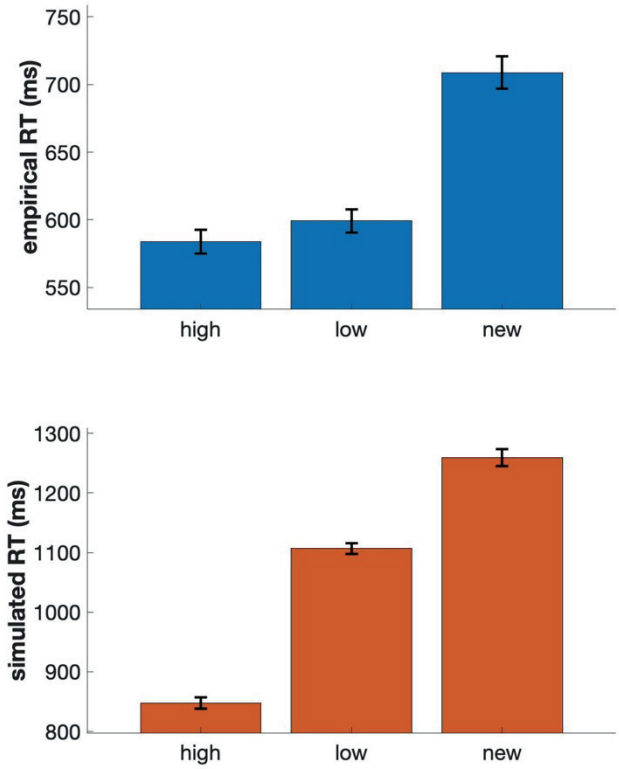

Figure 2. Lexical decision performance. High and low refer to 10-point words and 1-point words. New words were first seen in LD task.

Critically, figure 3 shows the simulated and empirical free recall data. Madan et al. (2012) show subjects recalled high-reward words often, followed by low-reward words, then new words. The simulated data reproduces this pattern extremely well. The simulated free recalls were also quantitatively accurate to the original empirical data ($R^2 = 0.85$, RMSE = 0.057). Overall, the model successfully reproduced the high-low relationship between reward and free recall reported in Madan et al. (2012).

Madan et al. (2012) also reported correlated free recall and lexical decision performance. Specifically, their analysis compared the number of recalled 10-point words normalized by the total number of recalls (i.e., proportion high-reward recalls) with a measure of LD facilitiation. The LD facilitation statistic was the difference in 10-point word RTs and 1-point word RTs, normalized by the subject's average RT. Madan et al. (2012) reported a negative correlation ($\rho(93) = -.2, p < .05$) indicating that, on average, subjects who recall high-reward words more often, also responded more quickly to high-value words. Figure 4 shows this relationship in the synthetic data as well.

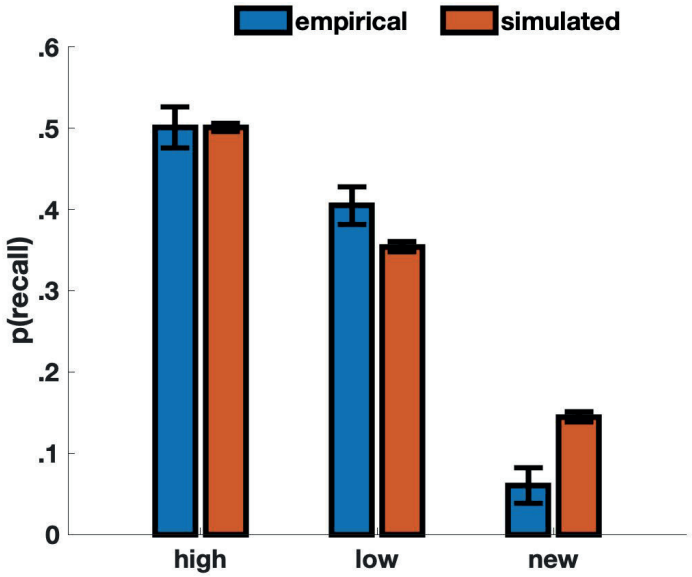

Figure 3. Free recall in Madan, Fujiwara, Gerson, and Caplan (2012) first experiment.

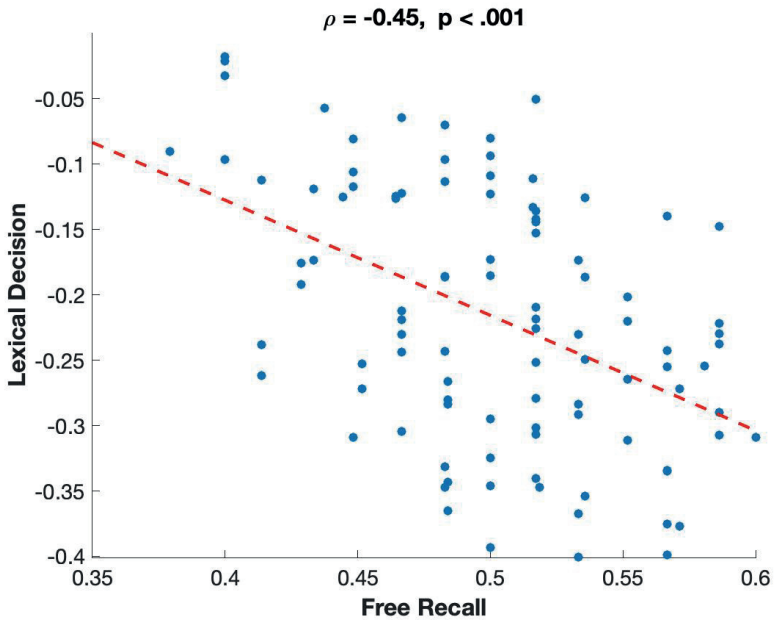

Figure 4. Correlation between free recall and lexical decision in synthetic data. See text for axes descriptions.

Notably, the simulated subjects' free recall performance varied due to noise in the recall process (specified by σ_D and θ_D), however subjects' LD performance did not meaningfully vary with all subjects sharing a fixed α_U value. Individual subjects all learned similar value functions, and the simple LD model included no noise term. Therefore, I added subject-wise variance in the reward expectation learning-rate parameter by drawing subject α_U values from $\sim N(\mu = .45, \sigma = .04)$. Varying subject learning rates introduced variance in their LD task performances needed for correlating that task performance with free recall.

Table 3. Goodness-of-fit for Madan et al. (2012).

	Lexical Decision		Free Recall		Intrusions	
Simulation	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Experiment 1	-66.728	458.380	0.850	0.057	-	-
Experiment 2	-	-	-14.453	0.035	0.976	0.007
Experiment 1, H_0	-0.253	62.349	0.955	0.038	-	-
Experiment 2, H_0	-	-	-17.675	0.042	0.351	0.025

2.2.2 Experiment 1 H_0 simulations

The same parameter search found appreciably different best-optimizing parameters for the null simulation (see Table 2). In this simulation the model does not include terms tying memory to value, representing an alternative to the full model including those terms specifying the hypothesis (also see section 2.1.8). Figure 5 shows the simulated value-learning reproduces the empirical behavior well (compare to Figure 1), this was expected since the terms modified for this simulation had no bearing on this task (9-10). Figure 6 shows the null simulation lexical-decision RTs (compare with Figure 2). Unlike the empirical data, simulated reaction times are equivalent for 10-point and 1-point words. The model does not qualitatively reproduce the results when RTs are untied from action-values (equations 5 and 9).

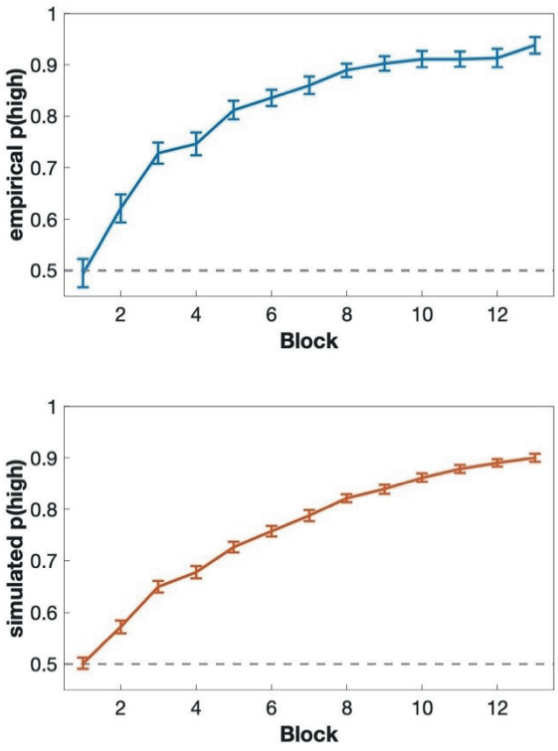

Figure 5. Value learning in first experiment H_0 simulation.

The goodness-of-fit statistics for these lexical-decision RTs (R^2 = -0.253, RMSE = 62.349) indicate the null simulation RTs fit the empirical data better than the full model (see Table 3). However, the R^2 statistics for both simulations are less than 0, indicating both simulations explain the data poorly (see Appendix for more information on negative R^2 statistics). The improvement in R^2 for the null model is less meaningful considering the statistic still shows a poor fit to the data. I argue the qualitative pattern of results is far more meaningful in this circumstance, especially considering the crude LD modeling. The R^2 statistics here do not reflect the null model's failure to show faster RTs to high reward items, a critical result.

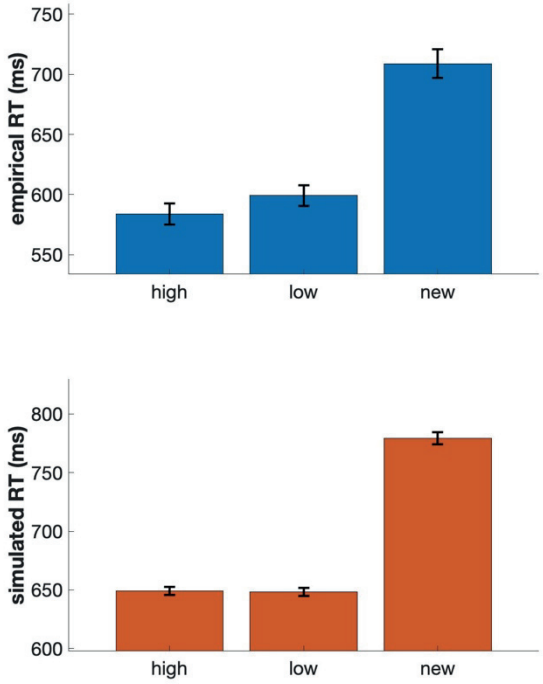

Figure 6. Lexical decision performance in first experiment H_0 simulation.

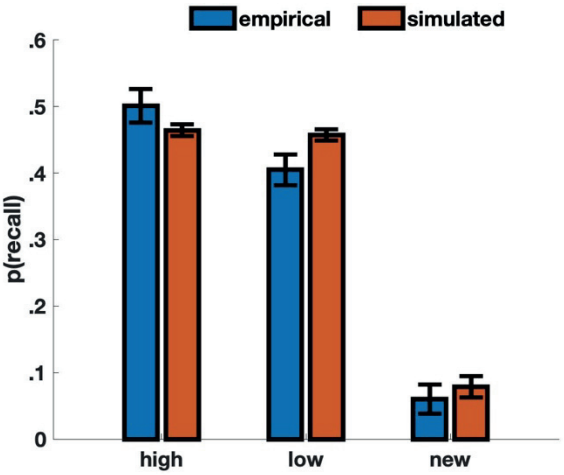

Figure 7. Free recall performance in first experiment H_0 simulation.

The free recall performance in figure 7 also does not replicate the empirical results (compare with figure 3). Subjects in the null simulation recalled high and low-reward words equivalently often. This indicates the model cannot reproduce the high-low relationship between value and memory strength without including terms specifying that hypothesis (9-10). Table 3 shows the free recall fit statistics actually improved slightly with the null model ($R^2 = 0.955$, RMSE = 0.038) and the R^2 values are closer to 1 (unlike the LD results). Again, I argue these statistics fail to reflect how the null model shows no memory differences between high and low reward items. The better memory for high-reward items is the most important empirical result in this experiment, and the null model cannot reproduce that result. Therefore, I argue the null model cannot explain the qualitative pattern of results as well, and that failure is more meaningful than improved goodness-of-fit (i.e., R^2). Furthermore, the correlation between LD performance and free recall (Figure 8) was positive, whereas the empirical correlation was negative (compare to figure 4). Overall, the null model was unable to show high-low reward effects on either implicit or explicit memory.

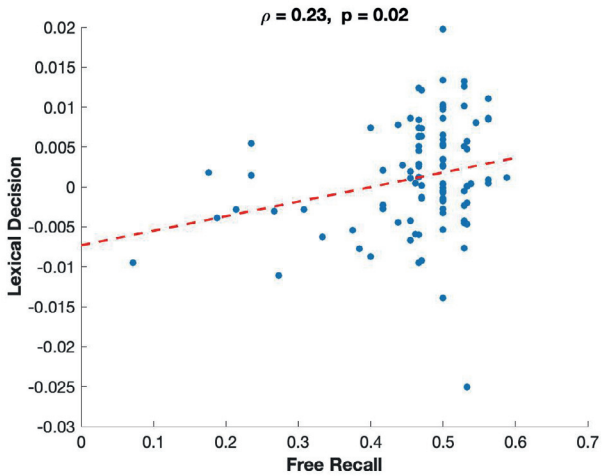

Figure 8. LD and free recall correlation in the H_0 simulation.

2.2.3 Simulations for Madan et al. (2012) Experiment 2

Subjects were simulated with the best parameter values found during parameter optimization (Table 2). The simulated value learning with adaptive α_W well-approximated the empirical data (figure 9). Free recall in the simulated study-free list cycles also matched the empirical results (R^2 = -14.453, RMSE = 0.035), as shown in figure 10. The experimental recalls for new words were numerically lower than simulated recalls, although Madan et al. (2012) reported no significant main effects for recall type (i.e., new and old recalls did not differ experimentally). The simulation accurately reproduced this pattern of results. Figure 11 shows the simulation also reproduced intrusion rates in the memory task extremely well (R^2 = 0.976, RMSE = 0.007). In both the experimental and synthetic data, subjects committed intrusion errors most often for 10-point words, less often for 1-point words, and least often for new words. This shows a high-low relationship between word value and intrusion rates in that more valuable words are harder to relearn in a new context. In other words, subjects' memories are strongest for high-value words and they therefore commit more intrusion errors for these stimuli.

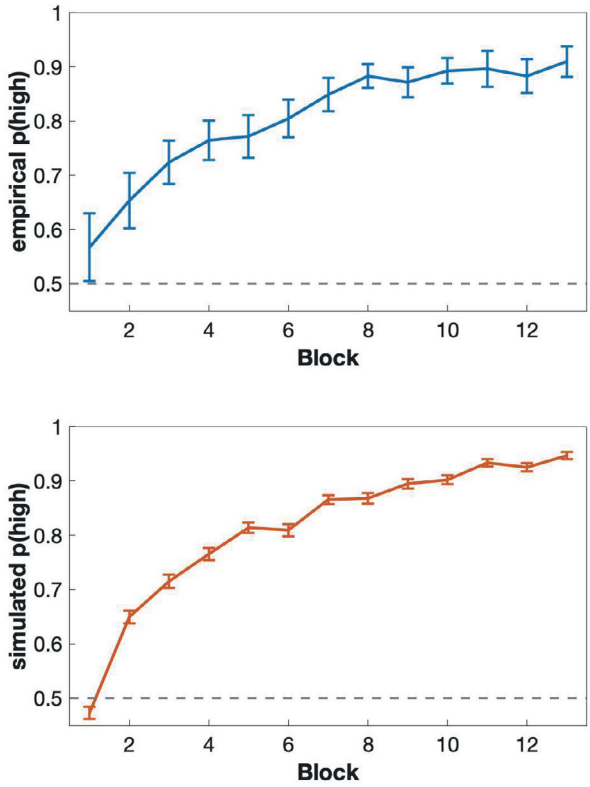

Figure 9. Value-learning in Madan et al. (2012) second experiment. Top panel shows experimental data, bottom panel shows simulation data.

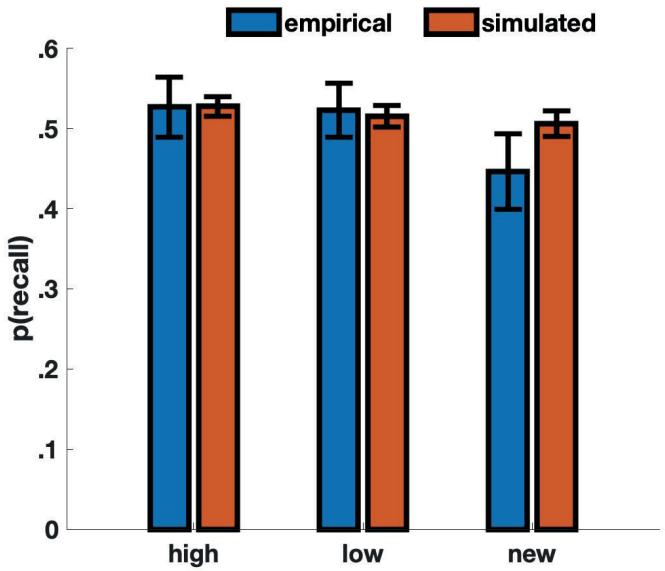

Figure 10. Free recall performance on study-recall cycles in Madan et al. (2012) second experiment.

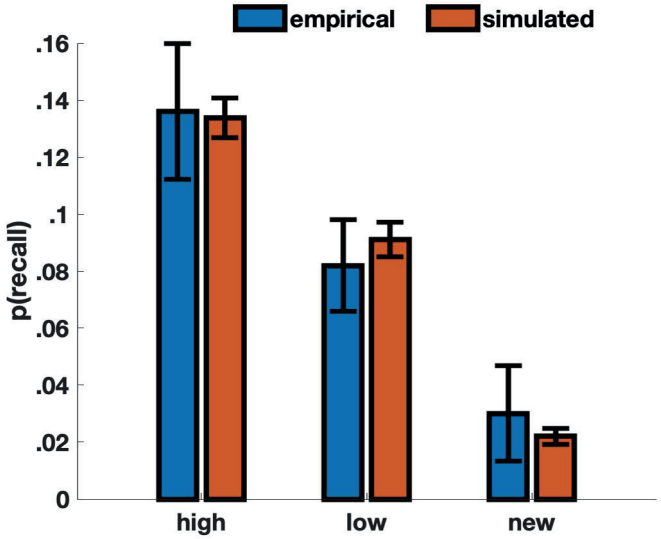

Figure 11. Intrusion rates during study-recall cycles in Madan et al. (2012) second experiment.

2.2.4 Experiment 2 H_0 Simulations

Table 2 shows the null model's best parameters found during optimization. Removing the terms tying memory strength to value had no impact on value-learning simulations (figure 12). Similarly, removing these terms did not prevent the model from reproducing experiment 2's correct free recall data (R^2 = -17.675, RMSE = 0.042; Figure 13). The experimental data showed no relationship between value and memory in correct recalls. Therefore, the null model was expected to reproduce the correct free recall data without terms relating value to memory.

Crucially, the null model was unable to reproduce intrusion rates in the free recall data (R^2 = 0.351, RMSE = 0.025; Figure 14). In the experimental data, subjects commit more intrusion errors for higher reward words. The null model's synthetic data shows that subjects only commit more intrusions for old words. Intrusion rates differ little for 10-point and 1-point words. If anything, the intrusion rates for low-words are slightly higher numerically, but that difference is probably not meaningful. The null model was unable to produce a relationship between value and intrusion errors.

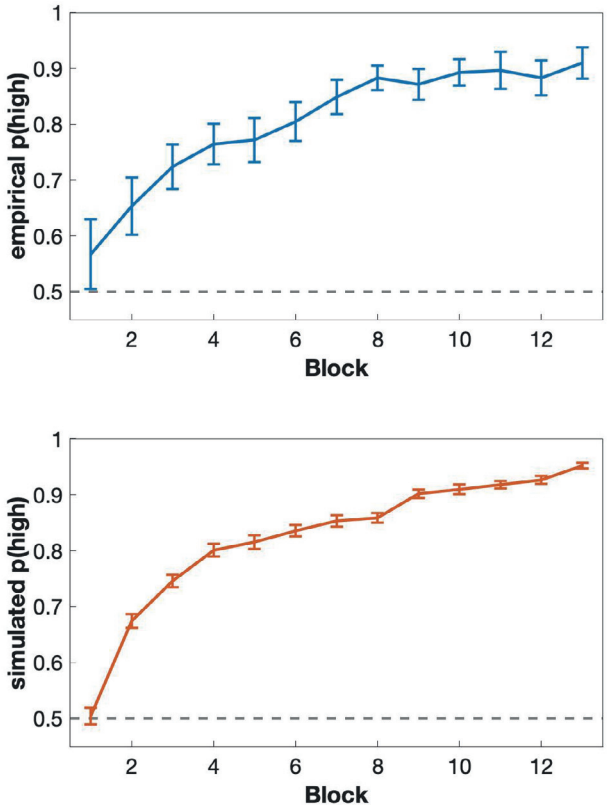

Figure 12. Value-learning in second experiment H_0 simulation.

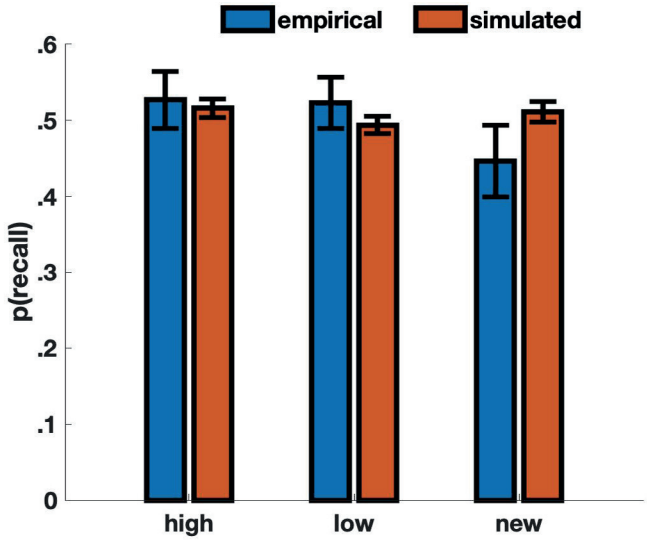

Figure 13. Free recall performance second experiment H_0 simulation.

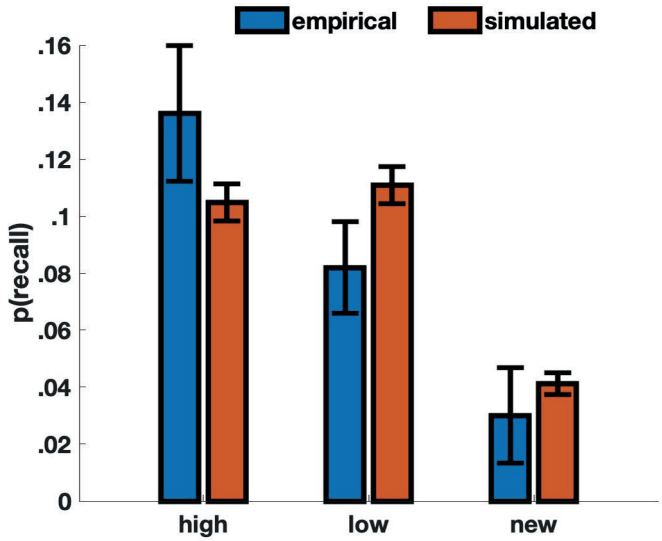

Figure 14. Intrusion rates during study-recall cycles in second experiment H_0 simulation.

3. U-shaped reward-memory outcomes

In Madan and Spetch (2012), subjects completed value-learning, lexical decision, and free recall tasks. This paradigm closely matched the experiments in Madan et al. (2012), except with multiple reward levels. Subjects' choices yielded three different point rewards (2, 7, or 12) in the first experiment, and six different rewards in the second experiment (2, 4, 6, 8, 10, or 12). Madan and Spetch (2012) found a U-shaped relationship between reward and memory, in which subjects best remembered both the highest-reward and lowest-reward items. Those extreme-reward items (i.e., highest and lowest) were remembered better than stimuli learned with intermediate or no rewards. Madan and Spetch (2012) concluded the U-shaped relationship was likely due to reward saliency, or parameter edge effects (e.g., the prototype memory effects in Huttenlocher, Hedges, and Duncan (1991)).

These results must connect with the high-low relationship observed previously in Madan et al. (2012) (aim 1), where subjects remembered high-reward stimuli better than low-reward stimuli. Madan et al. (2012) originally suggested attention or emotion may underlie their results, as attention and emotion have well-known influences on subsequent memory (as reviewed by Kensinger (2009)). Subjects may have paid more attention to high-reward items, or perhaps those items were more emotionally arousing. Madan and Spetch (2012) (aim 2) do note those explanations are not strictly incompatible with their later U-shaped data, if including more rewards produced some additional memory influence (e.g., an additive salience effect).

I hypothesized that memory strength depends on value, and that explains why higher reward yields better memory in Madan et al. (2012) (aim 1), but higher reward gives U-shaped memory outcomes in Madan and Spetch (2012) (aim 2). I predict the model will also capture the U-shaped relationship, because extreme-reward items are highly useful in 2-AFC tests. In this

context, recognizing an extremely low-reward target indicates that you should pick the alternative. Recognizing medium-reward targets does not offer such an advantage, as the alternative could be better or worse. Therefore, in Madan and Spetch (2012) both high-reward and low-reward items are valuable, and my model should explain the observed U-shape relationship with memory. This situation illustrates another instance where unrewarding experiences can be highly valuable. In the previous restaurant analogy, being seated at a table represents a highly valuable but unrewarded circumstance (i.e., the food is near, but there is no food present). Low-reward stimuli in 2AFC are more valuable because they indicate how the subject can obtain more reward.

3.1 Methods

3.1.1 Basic modeling description

I simulated the experiments in Madan and Spetch (2012) using the same modeling framework described in aim 1, with two important changes that are described here. First, I added a new set of action features to existing state-action representation. This new set contained features for rejecting possible actions. Borrowing the GO/NOGO task terminology (Gomez, Ratcliff, & Perea, 2007), these new features represent "NOGO" actions. The NOGO actions features for the i^{th} word are:

$$\overset{\emptyset}{a_i} = \begin{cases} 1, & s_i = 1 \; and \; a_i \neq 1 \\ 0, & otherwise \end{cases}$$

so that $\overset{\emptyset}{a_i}$ takes the value 1 when subject observes the i^{th} word, but does not choose the i^{th} word. That is, NOGO features represent the possible actions that are not taken. The total representation for a state-action pair is now

$$X(s,a) = \left\{1\,, \quad s_1 \ldots s_n\,, \quad s_{12} \ldots s_{\binom{n}{k=2}}\,, \quad a_1 \ldots a_n, \quad \overset{\emptyset}{a_i} \ldots \overset{\emptyset}{a_n}\right\}$$

The representation includes an intercept feature, a feature for observing each word, observing each combination of two words, choosing each word (GO actions), and not choosing each word (NOGO actions).

NOGO features are critical for modeling the hypothesis that recognizing low-reward stimuli offers an advantage in the learning task. With NOGO features the model explicitly learns and represents values for avoiding actions. Otherwise, the model's value function simply shows choosing lower-reward stimuli offers less value. The model implicitly learns to avoid these actions[3], but one would have to infer action-rejection values post hoc somehow. Including NOGO features provides the most straightforward way to model action-rejection values.

Secondly, I changed the memory evidence calculations. Equation 8 determined the memory strength in Madan et al. (2012) (aim 1) simulations. That implementation extended the Gershman et al. (2012) strength modeling to include relevant action features. Accordingly, Equation 8 was updated to include NOGO action features as well.

$$M_i = X_0^T W^T x\left(s_i, a_i, \overset{\emptyset}{a_i}\right) \tag{12}$$

Furthermore, the memory-evidence for each feature was scaled by the feature's value according to equation 4. This change was necessary because the empirical memory data in Madan and Spetch (2012) shows a U-shaped relationship between memory and reward; people remember the extreme-reward words best. However, the U-shape is slightly asymmetrical such that the highest-reward words are better remembered than the lowest-reward words. Madan and Spetch (2012) reported the asymmetry was statistically significant, and therefore the model should capture this effect. Scaling the feature-memory evidence (i.e., the memory strength for each feature, as per

[3] because the model's policy almost never chooses lower-valued options

equation 8) by value was included to produce this asymmetry. In full, these simulations calculated the memory evidence for each word *i* :

$$q = I_F WU \tag{13}$$

$$M_i = (X_0^T W^T \odot q)\, x\left(s_i, a_i, \overset{\emptyset}{a_i}\right) \tag{14}$$

where the vector q gives the value for each feature (as in equation 4), X_0 is a memory cue representation (just as in the previous equation 7).

3.1.2 Value learning task

The learning task simulations followed the same procedure previously described for Madan et al. (2012) (aim 1), except with the multiple reward levels described in Madan and Spetch (2012).

3.1.3 Lexical decision task

The lexical decision task simulations also followed the procedure described for Madan et al. (2012) (aim 1), except value estimations included NOGO features.

$$Q_i = x\left(s_i, a_i, \overset{\emptyset}{a_i}\right)^T WU \tag{15}$$

3.1.4 Free recall

Free recall simulations followed the same procedure described for Madan et al. (2012) (aim 1), with the updated memory strength equation 12.

3.1.5 Parameter optimization

Model parameters were optimized with the Matlab particle swarm optimizer *particleswarm(),* using 200 particles for all searches. In-house modifications allowed the optimizer to interface with a SLURM scheduler and run particle simulations in a cluster computing environment. The particle swarm optimized $a_U,\ a_W,\ \gamma,\ \lambda,\ \sigma_D$, and θ_D for both experiments. The objective scores for both experiments were:

$$\epsilon_{learn}{}^3 + \epsilon_{recall}{}^3$$

Lexical decision task statistics were not included in the objective score because Madan and Spetch (2012) found no relationship between reward and RTs, and subsequently the authors did not report RT summary statistics.

3.1.6 Simulations with H_0 models

I ran the same parameter search for two null hypothesis models; one model without scaling feature-memories by value, another model without NOGO features and without scaling feature-memories by value. The first null model (noted elsewhere as "H_0 #1") replaces memory strength equation 14 with:

$$M_i = X_0^T W^T \, x\left(s_i, a_i, \overset{\emptyset}{a_i}\right) \tag{16}$$

so feature-memories were not scaled by q (equation 13).

The second null model (noted elsewhere as "H_0 #2") also lacked NOGO features, so equation 14 was replaced with:

$$M_i = X_0^T W^T \, x(s_i, a_i) \tag{17}$$

The NOGO features were totally absent, and so they did not contribute to value learning either. Otherwise, these two null models were otherwise specified exactly as the primary model.

3.2 Results

3.2.1 Simulations for Madan and Spetch (2012) Experiment 1

Subjects were simulated with the best parameter values found during parameter optimization (Table 4). Figure 15 shows the simulated subjects' performance in the learning task. Simulated subjects reach performance levels comparable with the experimental data shown in Figure 16. In both simulated and empirical datasets, the average subject chooses correctly on > 70% of trials by the final block. Furthermore, both simulated and real subjects perform better

when words differ more in point associations. This reflects that choices between 2- and 12-point words are easier than choices between 7- and 12-point words (or 2- and 7-point words). However, the simulated choice behavior differs notably from the experimental data in the first few blocks. Simulated subjects reach nearly peak performance after one block, and improve little over the remaining blocks. In contrast, real subjects improve much more gradually. The slightly higher α_W (i.e., SR learning-rate) parameter value probably accounts for why learning improved suddenly in these simulations. Section 2.1.7 discusses the trade-off between high and low α values at length. In summary, high α enables fast learning at the expense of model volatility, whereas lower α gives slower learning and improves model stability. I set α_W and α_U at fixed values for the entire experiment for simplicity, and section 2.1.7 discusses that assumption's realism and implications. Here the larger, fixed α_W probably produced that qualitative mismatch between simulated and empirical learning curves. This mismatch represents a minor inaccuracy, with little bearing on the main hypothesis.

The model reproduced the U-shape memory results extremely well ($R^2 = 0.999$, RMSE = 0.004). Figure 17 shows the simulated free recall outcomes map almost perfectly onto the experimental data, reproducing the memory results qualitatively and quantitively. Simulated and real subjects best remembered 12-point words, followed by 2-point words, then 7-point words. Subjects also showed weak memory for words only encountered in the lexical decision task (i.e., new words).

Table 4. Best parameters found for Madan and Spetch (2012).

Simulation	α_U	α_W	γ	λ	σ_D	θ_D
Experiment 1	.291	.321	.579	.043	.01	1.241
Experiment 2	.397	.109	.665	.373	.021	1.591
Experiment 1, H_0 #1	.3	.055	.353	.498	.044	3.853
Experiment 1, H_0 #2	.384	.224	.016	.308	.012	3.146
Experiment 2, H_0 #1	.3	.055	.353	.498	.044	3.853
Experiment 2, H_0 #2	.267	.058	.379	.526	.045	3.894

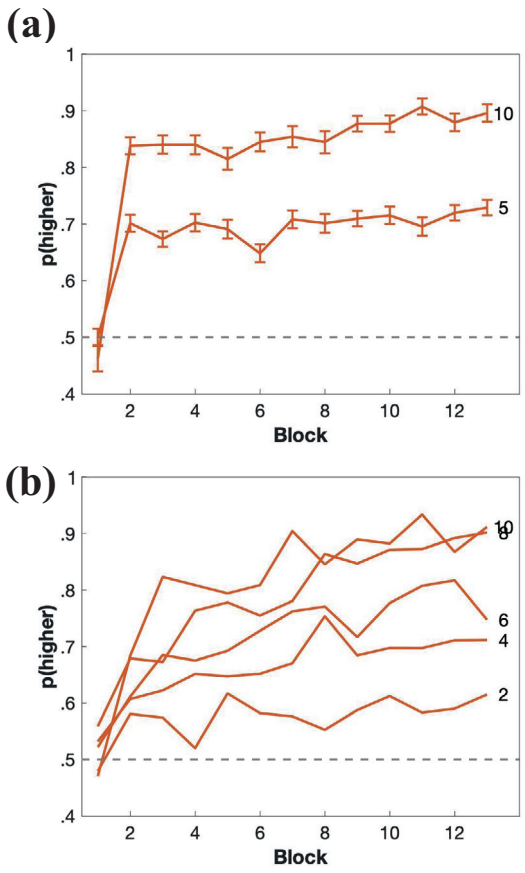

Figure 15. Probability higher-reward word was chosen, plotted by difference in item reward. Simulated Madan and Spetch (2012) learning task, (a) 1st and (b) 2nd experiment.

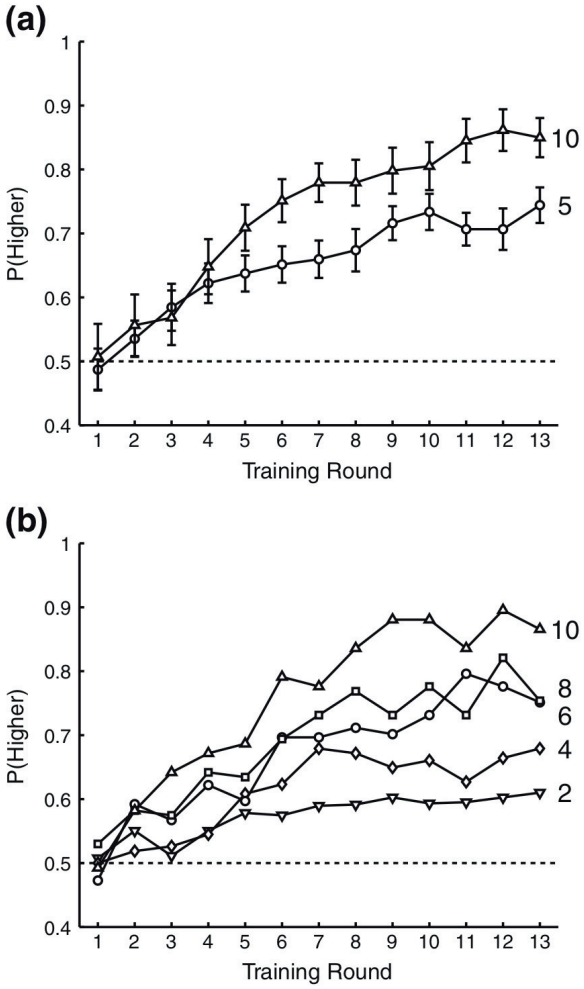

Figure 16. Reprinted from Madan and Spetch (2012). Empirical value learning behavior (a) first experiment and (b) second experiment.

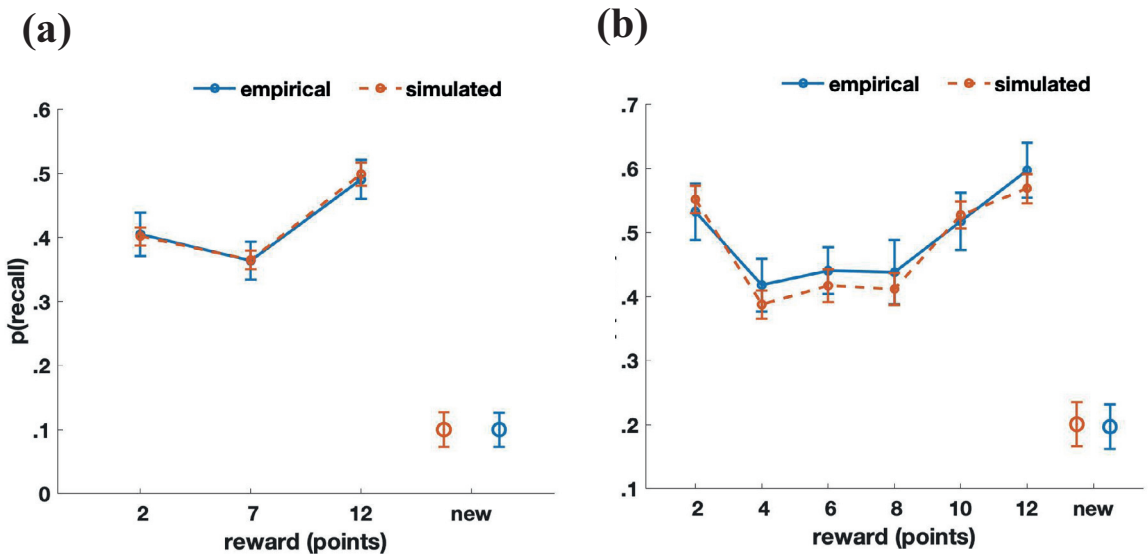

Figure 17. Free recall in Madan and Spetch (2012) first experiment (a) and second experiment (b)

3.2.2 Simulations for Madan and Spetch (2012) Experiment 2

Table 4 shows the best parameter values found during optimization, and those values were used for subject simulations. Figure 15 shows the simulated subjects' performance in the learning task (compare with figure 16). Simulated subjects' peak performance was comparable to the experimental data. In both simulated and empirical datasets, the average subject chooses correctly on > 85% trials with a 10-point word difference (i.e., the easiest trials) by the final block. The simulated subjects appear to improve faster than real subjects, but the difference is much less pronounced compared to experiment 1's results. The smaller α_W value and better matching (i.e., more gradual) learning curves also indicates the high α_W is probably why learning improved so dramatically in the first experiment's simulations.

Figure 17 shows the model reproduced the U-shaped reward-memory relationship very well ($R^2 = 0.965$, RMSE = 0.022). Simulated subjects remembered extreme-reward items (i.e., 2- and 12-point words) better than intermediate reward items. The free recall also showed the asymmetry seen in the empirical data, such that subjects also show a relative memory advantage for 10-point words. In addition to capturing the results qualitatively, the model was also quantitively accurate.

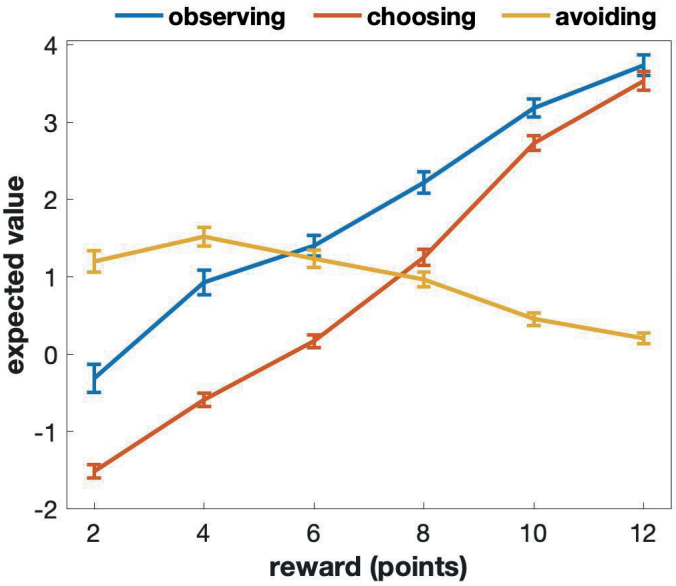

Figure 18. Value function after choice task simulations (experiment 2). Y axis shows value estimate for features: observing words X(s), choosing words X(a), avoiding words X(a-NOGO).

Figure 18 shows the model value functions immediately following the choice task simulations (section 3.1.2). The model expects greater value for both observing and choosing higher-point words, just as one would anticipate. The model also successfully learned to expect more value for avoiding low-point words. These curves help illustrate how the model reproduces the U-shaped memory results in Madan and Spetch (2012) with the value function terms in equation 14.

Table 5. Goodness-of-fit for Madan and Spetch (2012).

	Free Recall	
Simulation	**R^2**	**RMSE**
Experiment 1	0.999	0.004
Experiment 2	0.965	0.022
Experiment 1, H_0 #1	0.968	0.023
Experiment 1, H_0 #2	0.623	0.085
Experiment 2, H_0 #1	0.772	0.048
Experiment 2, H_0 #2	0.839	0.051

3.2.3 Simulations with H_0 models

Table 4 shows the optimal parameters found for both null models (i.e., H_0 #1, H_0 #2) in experiments 1 and 2. The results in this section were obtained with those parameters. Figure 19 shows learning performance in both experiments. The learning curves closely resemble the empirical curves shown in figure 16. As expected, the null models still capture learning behavior well. These simulations also show more gradual performance improvements with lower α_W values compared with the simulation 1 results. That pattern was also observed in the experiment 2 results, and further indicates the rapid learning improvement in experiment 1 simulations was simply due to higher α_W. Otherwise, the performance in figure 19 closely resembles performance shown in figure 15 (full model) and figure 16 (empirical data), without remarkable differences.

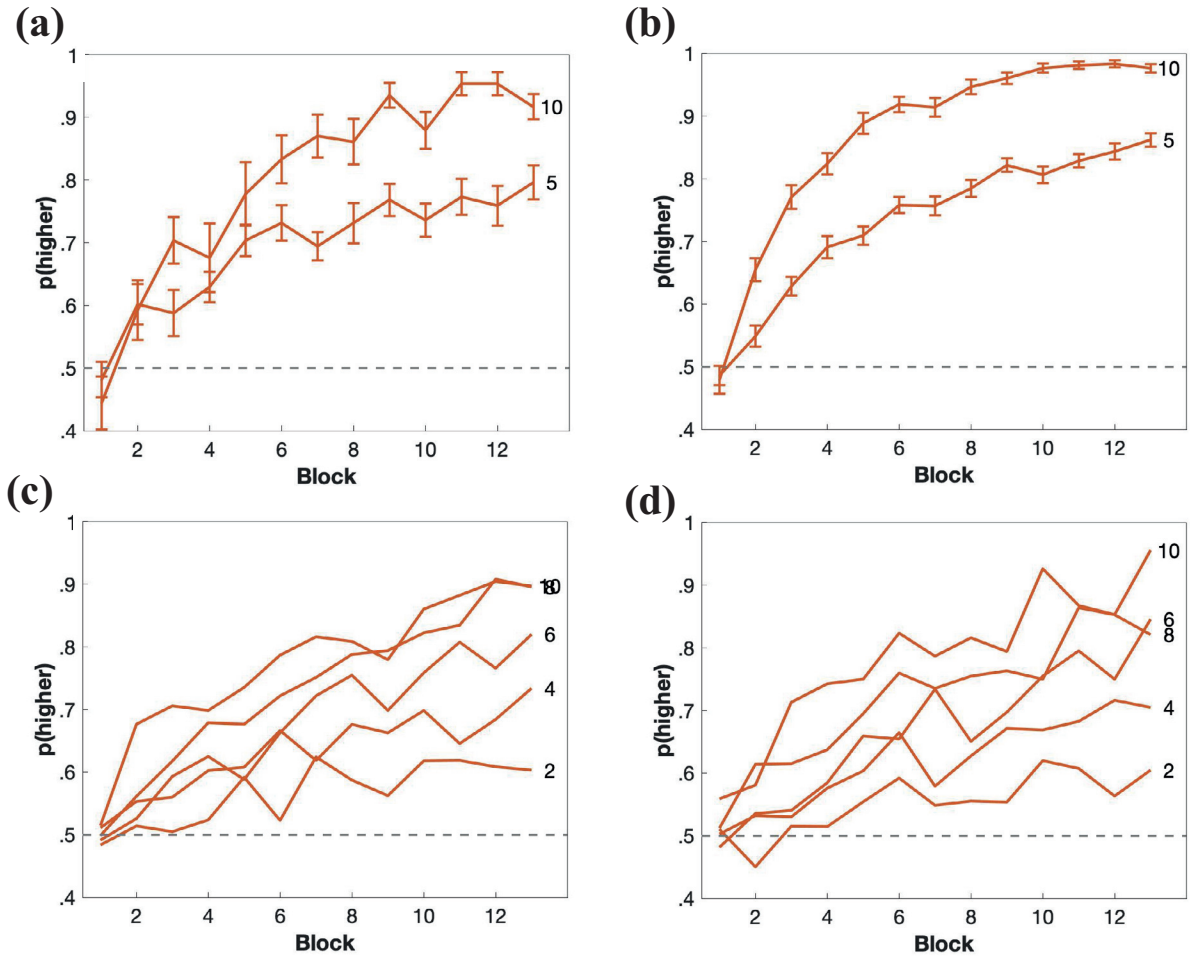

Figure 19. Value learning in Madan and Spetch (2012) H_0 Simulations. First experiment, H_0 #1 (a); first experiment, H_0 #2 (b); second experiment, H_0 #1 (c) second experiment, H_0 #2 (d).

The null models' simulated free recall performance in both experiments are shown in Figure 20 (see section 3.1.6 for methods descriptions). Panel (a) shows experiment 1 simulations without value scaling (equation 16), and the null model fails to reproduce the experimentally observed data. Panel (b) shows free recall in experiment 1 without value scaling or NOGO features (equation 17). This model also fails to reproduce the U-shaped memory results, as subjects recalled 2- and 7-points words equivalently often. The experimental data shows subjects

better remembered 2-point words compared to 7-point words, and the model was unable to reproduce that.

Panels (a) and (b) show simulated free recall in the second experiment, for the H_0 #1 and H_0 #2 models respectively. Neither model produces the U-shaped memory result observed experimentally. The models that do not scale memory by value, and do not include NOGO features (i.e., H_0 #1 and H_0 #2) are unable to reproduce the U-shaped memory results in Madan and Spetch (2012). Table 5 shows goodness-of-fit statistics for these simulations alongside those for the full model.

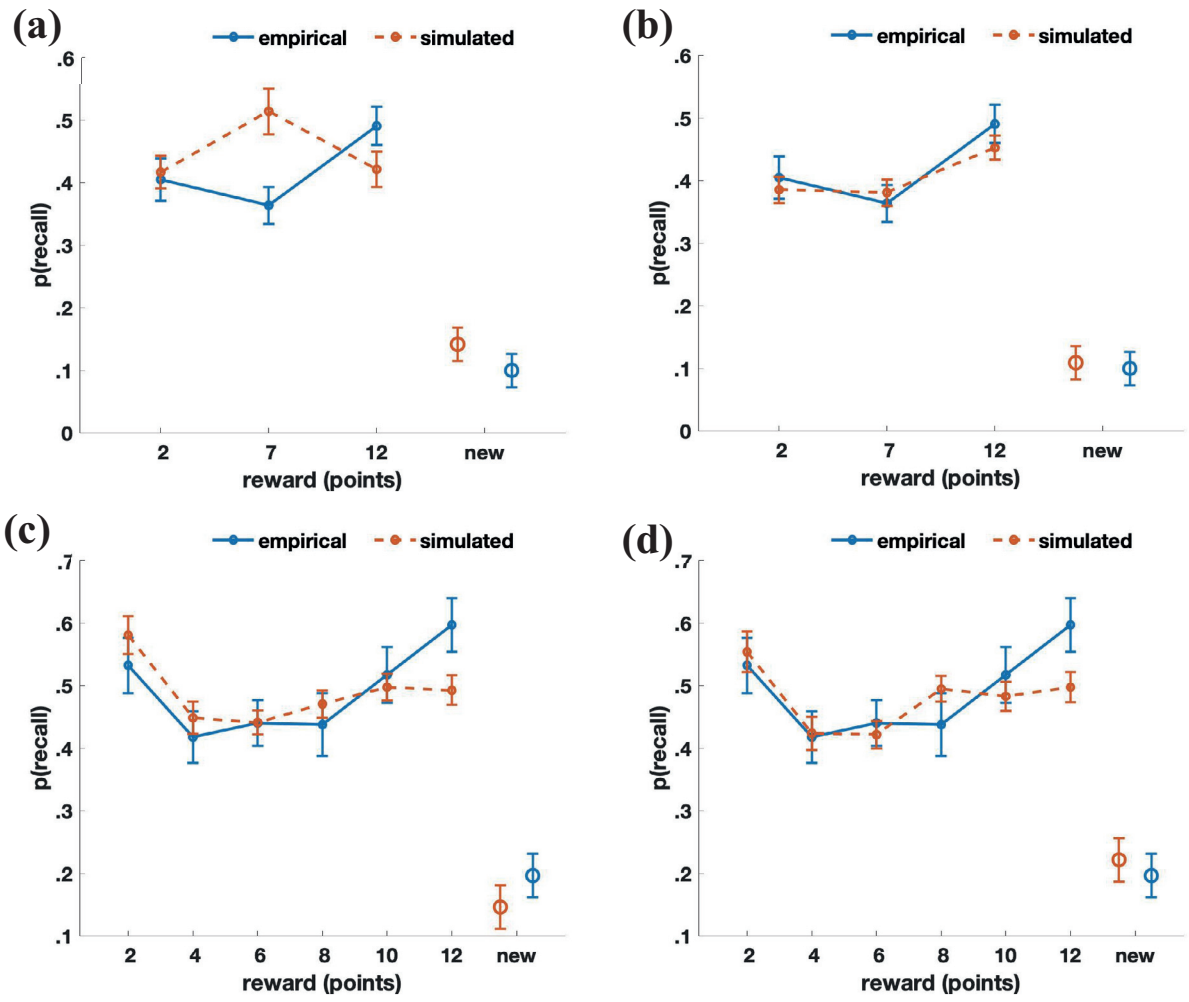

Figure 20. Free recall in Madan and Spetch (2012) H_0 Simulations. First experiment, H_0 #1 (a); first experiment, H_0 #2 (b); second experiment, H_0 #1 (c) second experiment, H_0 #2 (d).

4. Attenuated reward-memory outcomes

Chakravarty et al. (2019) employed a 2-AFC learning task where subjects chose between trial-unique words and the fixed letter string "HHHHH" (Figure 21). In this task subjects can earn the maximum possible reward on each trial. Furthermore, subjects learn through sequential experiences with only one word stimulus at a time. This procedure directly contrasts with value learning in Madan et al. (2012) and Madan and Spetch (2012) in which subjects learn from experiences involving two word stimuli simultaneously. Chakravarty et al. (2019) describes the latter scenario as learning "in competition", and I will use that terminology as well. The authors also describe the word-types as "high-value" and "low-value" (Figure 21). I will refer to these conditions as "10-point" and "1-point" words, respectively. This terminology avoids conflating value with reward.

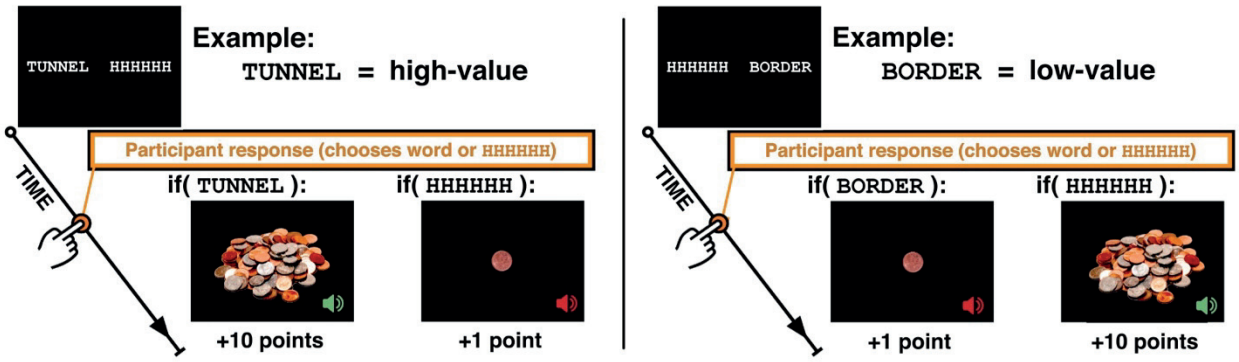

Figure 21. Value learning task. Reprinted from Chakravarty et al. (2019Journal of Memory and Language), Journal of Memory and Language.

In three experiments Chakravarty et al. (2019) demonstrates this value learning procedure attenuates the reward-memory relationship with subsequent free recall tests. That is, subjects do not remember words in the 10-point condition (Figure 21) differently from words in the 1-point condition. In the first experiment subjects completed the value learning task, followed by a

lexical decision task, and finally free recall. This protocol followed the same structure as experiments in Madan et al. (2012) (aim 1) and Madan and Spetch (2012) (aim 2).

The second experiment counterbalanced the LD and free recall ordering across subjects (i.e., half the subjects performed free recall before LD), and yielded rewards probabilistically instead of deterministically. Specifically, the deterministic outcomes were replaced with 9:1 outcome odds. The optimal response yielded 10-points 90% of the time, but only 1-point 10% of the time. Conversely, the suboptimal response gave 1-point on 90% of the time, and 10-points 10% of the time.

The third experiment did not include a lexical decision-task, subjects only completed value-learning (with probabilistic rewards) followed by free recall. Subjects were also rewarded for their free recall performance in this experiment, and subjects were assigned to one of three free recall conditions. In the congruent conditions, subjects received 300 points for recalling 10-point words (i.e., "high value") and 100 points for 1-point (i.e., "low value") words. In the incongruent condition, subjects received the opposite; 100 points for 10-point and 300 points for 1-point words. In the neutral condition, subjects received 200 points for each correct recall regardless of word-type.

Subjects showed an attenuated relationship between reward and memory in all three experiments. Chakravarty et al. (2019) reported their results converged on null effects, such that subjects remembered 1-point and 10-point words equally often. I hypothesized that memory strength depends on value, such that people should remember more valuable information easier. The unique learning paradigm in Chakravarty et al. (2019) avoids learning in competition, but also effectively equates the value for all stimuli in the experiment. Subjects can earn the maximum reward on every trial, so the well-trained subject should expect the maximum value

for each stimulus and its correct corresponding action (i.e., either choosing or avoiding). Therefore, well-trained subjects should value all stimuli equally, and remember them all equally well. The model should both reproduce all the null memory results, and the value function should correspondingly show little difference between 1-point and 10-point words.

4.1 Methods

4.1.1 Basic modeling description

I simulated the experiments in Chakravarty et al. (2019) using the same modeling framework described in aim 2. The model was not modified further for these simulations. Furthermore, the task simulations were also fundamentally unchanged from their implementations in aims 1 and 2.

4.1.2 Value learning task

The learning task simulations followed the same procedure previously described for Madan and Spetch (2012) (aim 2), except with probabilistic rewards for experiments 1 and 2. The probabilistic rewards assigned the associated point value 90% of the time, and the alternate reward 10% of the time. For instance, choosing a 10-point word (see Figure 21) yielded a 10-point reward with $p = .9$, and 1 points with $p = .1$.

4.1.3 Lexical decision task

The lexical decision task simulations also followed the procedure described for Madan and Spetch (2012) (aim 2).

4.1.4 Free recall

Free recall simulations followed the same procedure described for Madan and Spetch (2012) (aim 2). The rewarded recall in experiment 3 still follows the same implementation described in the section 2.1.6 pseudocode, except reward was not fixed at zero. Subjects in the

congruent condition received 300 points for recalling 10-point words, and 100 points for 1-point words. Subjects in the incongruent condition received 100 points for 10-point and 300 points for 1-point words. Subjects in the neutral condition received 200 points for each correct recall regardless of word-type.

In experiment 2, half the subjects complete free recall before the lexical-decision task (FR-LD condition) and the other half complete LD first (LD-FR condition). Both conditions were simulated, but the results were aggregated as Chakravarty et al. (2019) reported non-significant differences between FR-LD and LD-FR conditions.

4.1.5 Parameter optimization

Model parameters were optimized with the Matlab particle swarm optimizer *particleswarm(),* using 200 particles for all searches. In-house modifications allowed the optimizer to interface with a SLURM scheduler and run particle simulations in a cluster computing environment. The particle swarm optimized $a_U,\ a_W,\ \gamma,\ \lambda,\ \sigma_D$, and θ_D for all three experiments. The objective scores for experiments 1 and 2 were: ${\epsilon_{learn}}^3 + {\epsilon_{LD}}^3 + {\epsilon_{recall}}^3$. Experiment 3 did not include lexical decision, so the objective score was simply: ${\epsilon_{learn}}^3 + {\epsilon_{recall}}^3$.

4.1.6 Simulations with H_0 models

I ran the same parameter search for three null hypothesis models; one model without scaling feature-memories by value, a second model without NOGO features, and a third model without both NOGO features and without scaling feature-memories by value. The first null model (noted elsewhere as "H_0 #1") replaces memory strength equation 14 with:

$$M_i = X_0^T W^T\, x\left(s_i, a_i, \overset{\emptyset}{a_i}\right) \tag{18}$$

so that feature-memories were not scaled by q (equation 14).

The second null model (noted elsewhere as "H_0 #2") lacked NOGO features, but included feature-memories scaled by value. This model replaced equation 14 with:

$$M_i = (X_0^T W^T \odot q)\, x(s_i, a_i) \tag{19}$$

The third null model (noted elsewhere as "H_0 #3") lacked both NOGO features and value scaled memory evidence, so equation 14 was replaced with:

$$M_i = X_0^T W^T\, x(s_i, a_i) \tag{20}$$

The NOGO features were totally absent, and so they did not contribute to value learning either. Otherwise, these three null models were otherwise specified exactly as the primary model. I simulated experiments 1 and 3 with these null models, but not experiment 2. The second experiment simply swapped the lexical decision and free recall task ordering for one group of subjects. Therefore, simulating experiment 2 with the null models was unlikely to provide any more insight beyond the experiment 1 null simulation results.

4.2 Results

4.2.1 Value learning

Subjects were simulated with the best parameter values found during parameter optimization (Table 6). Figure 21 shows the simulated subjects' performance in the learning task. Simulated subjects reach performance levels comparable with the experimental data shown in Figures 23 and 24. In both simulated and empirical datasets, the average subject chooses correctly on > 85% of trials by the final block in experiment 1. Peak performance decreases in experiments 2 and 3 in response to probabilistic rewards, as expected.

Table 6. Best parameters found for Chakravarty et al. (2019).

Simulation	α_U	α_W	γ	λ	σ_D	θ_D
Experiment 1	.274	.198	.103	.264	.271	6.918
Experiment 1, H_0 #1	.229	.134	.273	.4	1.286	30.438
Experiment 1, H_0 #2	.194	.281	.356	.551	.397	9.849
Experiment 1, H_0 #3	.18	.281	.31	.563	.475	12.034
Experiment 2	.054	.25	.684	.352	.682	15.99
Experiment 3	.164	.267	0	.906	.014	3.061
Experiment 3, H_0 #1	.157	.24	.049	.240	.011	.1

4.2.2 Lexical decision

Figure 25 shows both the simulated and empirical lexical decision data in experiments 1 and 2. The model captures the RT pattern for high (10-point) and low (1-point) words, such reaction times did not differ between reward levels. However, the model does not reproduce the basic old/new priming result observed in the experimental data. Chakravarty et al. (2019) does not report statistical tests comparing RTs for new and old words, although the mean difference is at least 50ms in both experiments. The model could not produce an RT difference between old and new in that magnitude, likely due to negligible difference in word values (equation 14).

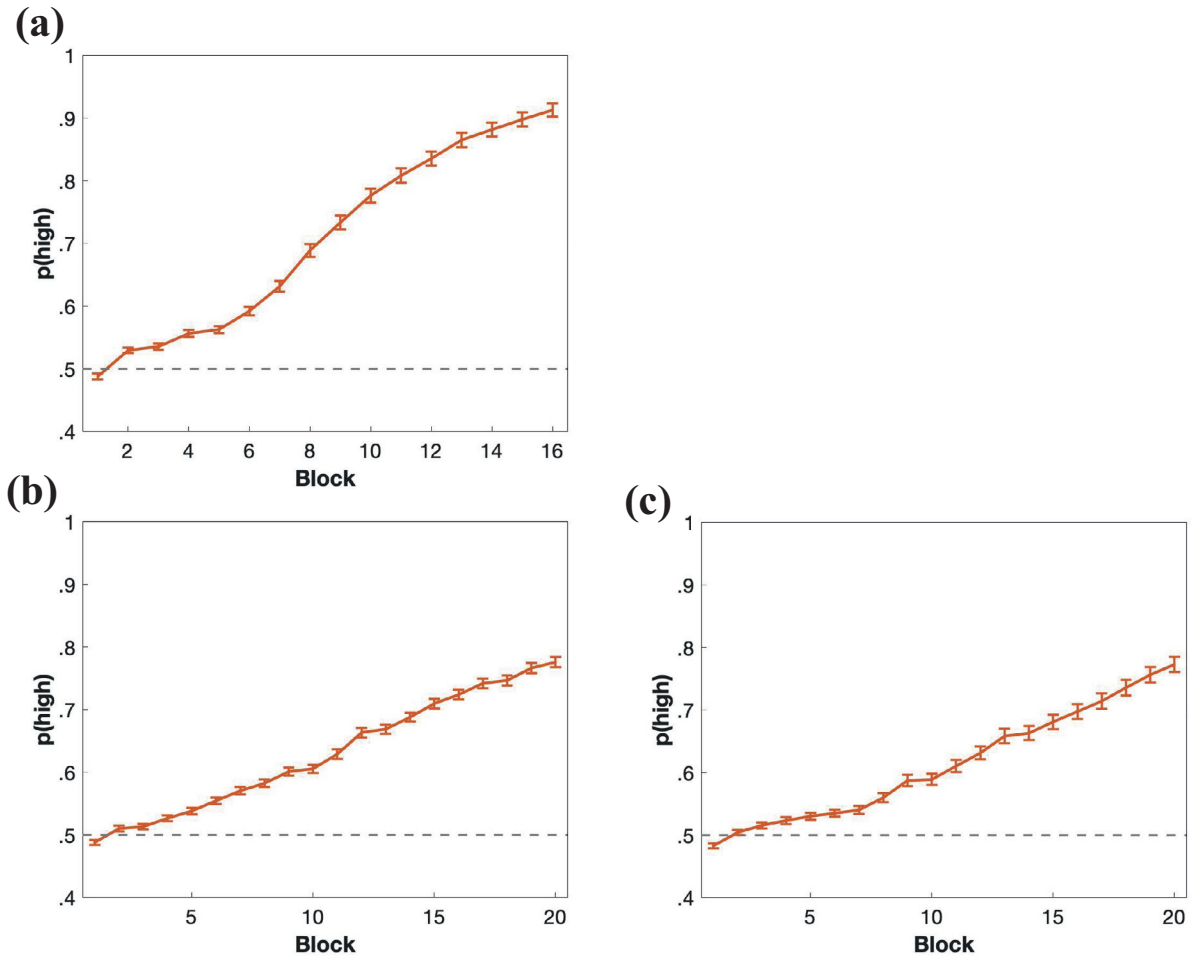

Figure 22. Value learning in Chakravarty et al. simulations: experiment 1 (a), experiment 2 (b) and experiment 3 (c). Y axes show probability correct option was chosen.

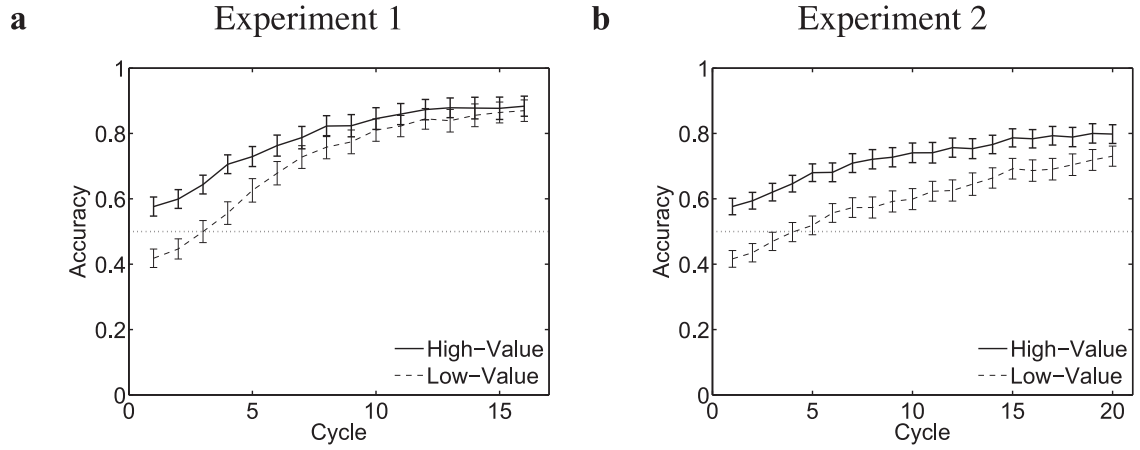

Figure 23. Reprinted from Chakravarty et al. (2019). Empirical value learning behavior (a) first experiment and (b) second experiment.

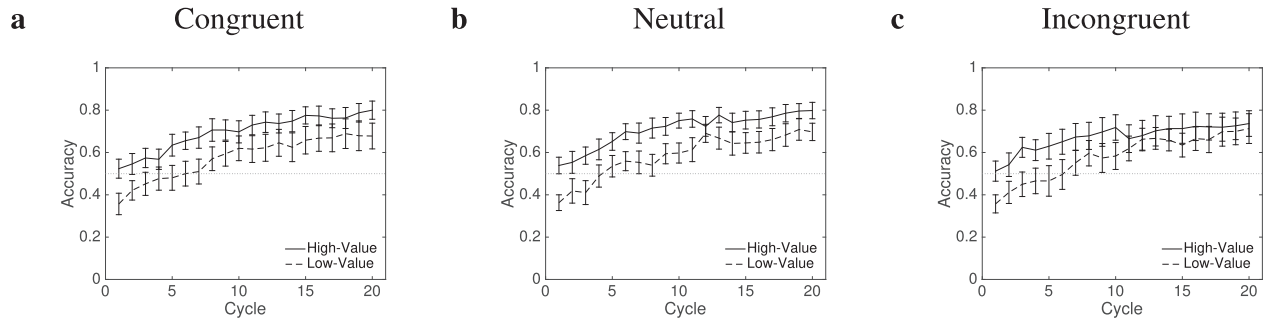

Figure 24. Reprinted from Chakravarty et al. (2019). Empirical value learning behavior in the third experiment, panels show individual conditions.

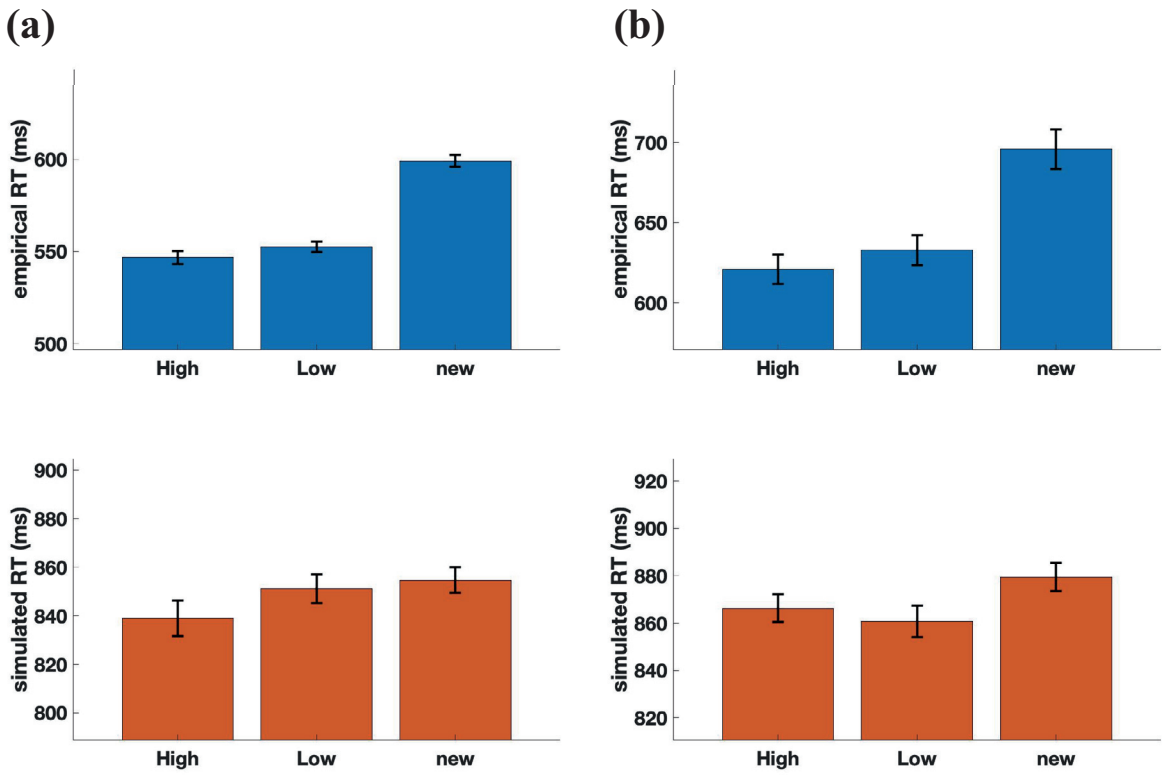

Figure 25. Lexical decision performance in Chakravarty et al. experiment 1 (a), and experiment 2 (b).

4.2.3 Free recall

Figure 26 compares the simulated and empirical data in the first two experiments. The empirical data shows no difference between high and low words in either experiment. The model reproduces that pattern with both qualitative and quantitative accuracy. The model succeeds because word memory-strength comprises the memory for seeing the word, choosing the word, and avoiding the word (14). The value function also scales feature memory strength, and Figure 27 shows that function immediately after simulating value learning.

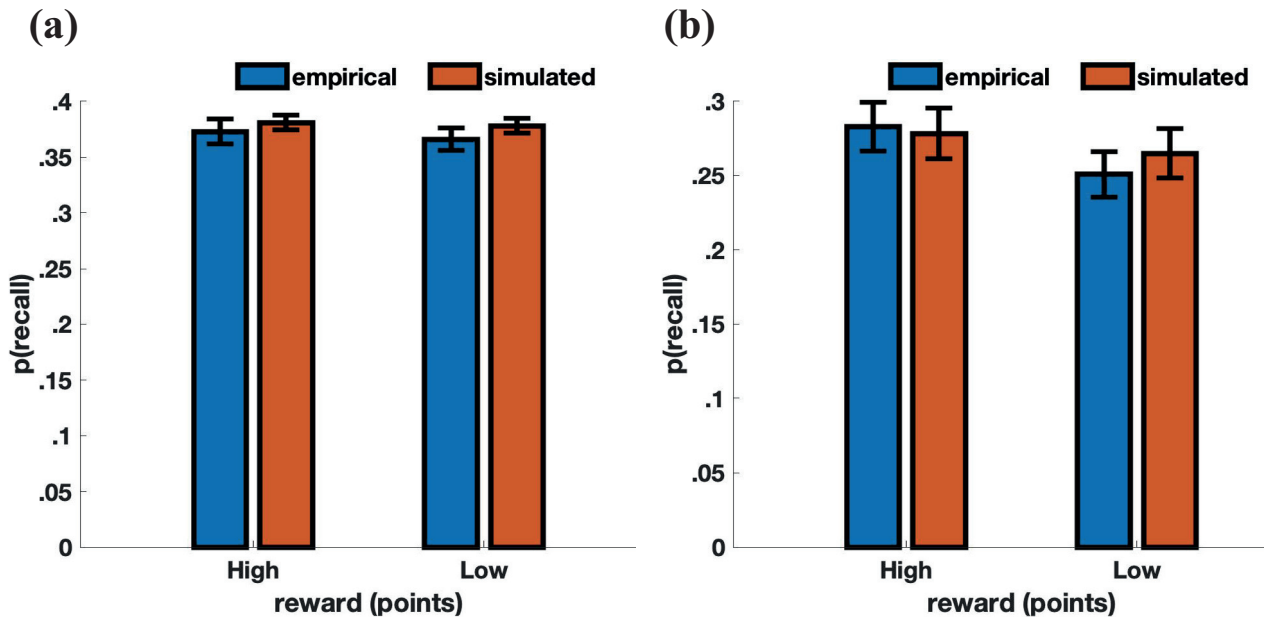

Figure 26. Free recall in Chakravarty et al. experiment 1 (a), and experiment 2 (b).

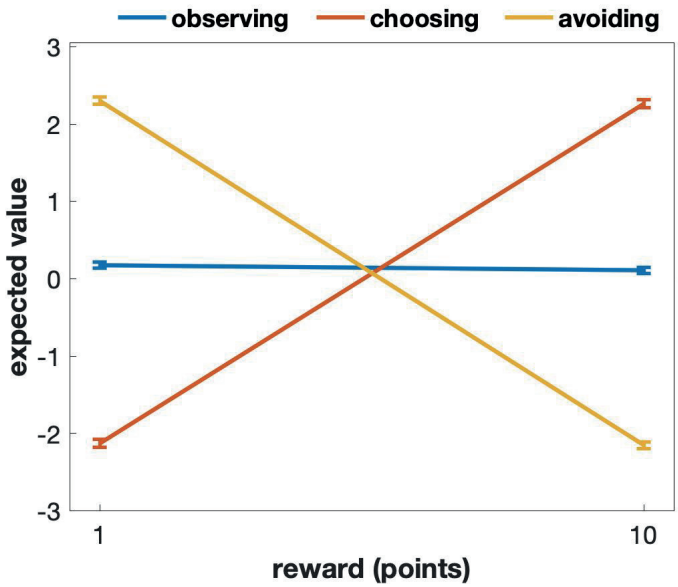

Figure 27. Value function after choice task simulations (experiment 1). Y axis shows value estimate for features: observing words X(s), choosing words X(a), avoiding words X(a-NOGO).

Figure 27 shows the model anticipates the same value for observing 1- and 10-point words, as expected. Similarly, the expected values for avoiding a 1-point word (i.e., choosing "HHHHH") and choosing a 10-point word are equivalent. Therefore, all words have the same cumulative value for observing, choosing, and avoiding the word. This makes the value biases in equation 14 irrelevant to memory outcomes. This does not interfere with learning in the 2AFC task because decisions are essentially comparing two state-action pairs: observing and choosing the target word, or observing and avoiding the target word. The subject cannot possibly both choose and avoid a word simultaneously. However, all features relevant to a word contribute to recall in this model.

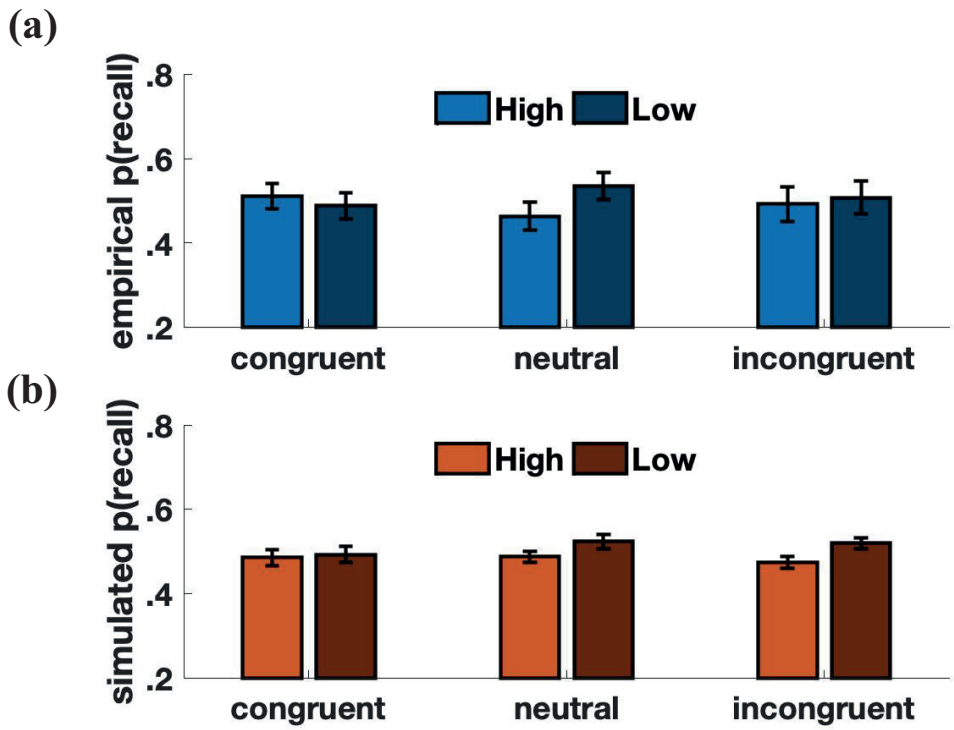

Figure 28. Free recall in Chakravarty et al. experiment 3, experimental (a) and simulated data (b). See text for condition descriptions (x axes).

Experiment 3 rewarded free recall, and the rewards followed one of three conditions. Subjects in the congruent condition were given higher rewards for recalling 10-point words. Subjects in the neutral condition were rewarded equally for all correct recalls. Finally, subjects in the incongruent condition were given higher rewards for remembering 1-point words (i.e., the high/low reward associations reversed). Figure 28 compares the experimental and simulated data across all three conditions. Chakravarty et al. (2019) reports no differences in memory in any condition; subjects remembered 1- and 10-point words equally well in all conditions. The model reproduces that pattern qualitatively, and with little numerical inaccuracy. Overall, the model captures the attenuated reward-memory relationship extremely well across all three simulated experiments.

Table 7. Goodness-of-fit for Chakravarty et al. (2019).

	Lexical Decision		Free Recall	
Simulation	**R^2**	**RMSE**	**R^2**	**RMSE**
Experiment 1	-143.750	282.770	-50.037	0.010
Experiment 1, H_0 #1	-43.878	220.410	-1.567	0.011
Experiment 1, H_0 #2	-	-	0.052	0.018
Experiment 1, H_0 #3	-173.600	310.560	-174.700	0.006
Experiment 2	-172.250	309.350	-9.032	0.011
Experiment 3	-211.910	342.940	-11.265	0.024
Experiment 3, H_0 #1	-	-	-11.830	0.124

4.2.4 Simulations with H_0 models

Table 4 shows the best parameters found when simulating experiments in Chakravarty et al. (2019) with the null models. In experiment 1 simulations, all three null models yielded unremarkable results (not shown for brevity). All three null models reproduced the attenuated free recall results equally well, with comparable accuracy to the primary model. The authors in Chakravarty et al. (2019) deliberately removed the reward-memory relationship through changing the value learning task. Therefore, the null models are more likely to reproduce the experimental data in this situation because the data lacked an experimental effect. The models simply needed to show subjects remembered words equally well when they studied those words an equal number of times. This is a trivial result for any successful memory model.

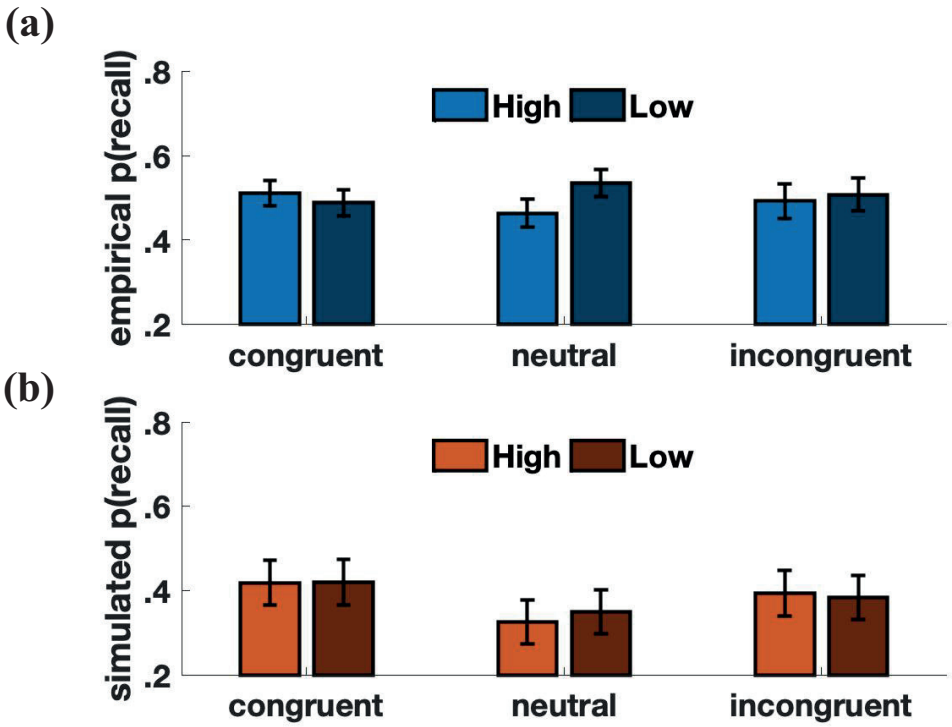

Figure 29. Free recall in Chakravarty et al. experiment 3, model without value scaling feature-memories (H_0 #1).

However, removing the value-scaling term (H_0 model #1) from experiment 3 simulations does noticeably degrade free recall numerical accuracy. Figure 29 shows these results with the experimental data (compare with Figure 28). The null model still reproduces the data qualitatively; subjects remember high and low words equally well. However, the model's numerical fit suffers without scaling feature-memory evidence by value. Table 7 shows the RMSE decreases slightly from 0.011 to 0.124 when comparing the full and null model's goodness of fit statistics for experiment 3. Combined with simulations in aim 2 showing this term is necessary for reproducing the effects in Madan and Spetch (2012), Figure 29 gives some tentative evidence that a value memory bias is involved in the Chakravarty et al. (2019) results. That is notable because the authors in Chakravarty et al. (2019) hypothesized behavioral

differences drive reward memory effects (i.e., choosing one word over another), whereas the value-scaling term represents an implicit value bias in memory.

5. Predictions for restoring attenuated outcomes

Chakravarty et al. (2019) proposes that learning item values in competition produces reward-memory effects, therefore learning item values one at a time attenuates those effects. That is, reward should only influence memory when subjects learn values via comparisons, or when they must learn about multiple items simultaneously. However, the current study's model should predict reward-memory effects outside those situations because the value function scales memory strength at retrieval.

An interesting aspect of the Chakravarty et al. (2019) learning task is that subjects can win the maximum reward on each trial. Successful models need to learn which words to choose and avoid, but these actions yield the same cumulative reward for each word (figure 27). Avoiding 1-point words yields the same reward as choosing 10-point words, and vice versa. Therefore, encountering any word signals the same expected value. The current study's model biases memory strength at retrieval according to equations 13 and 14; higher expected value increases memory strength. In this unique situation, however, that bias should have no meaningful impact because the words are uniformly valuable. This leads to the basic prediction that non-uniform values will restore reward-memory effects in the Chakravarty et al. (2019) task.

I evaluated this idea by simulating the Chakravarty et al. (2019) learning task with different reward contingencies. In these simulations, subjects learned item values one at a time, but certain trials could yield more reward than others. I hypothesized the model should predict reward-memory effects in this situation. If changing the value function restores reward-memory effects, those novel experimental predications can be tested empirically. This provides an

opportunity for falsifying the model, and perhaps informs whether reward influences memory beyond situations with competitive learning.

5.1 Methods

In these simulations, the learning task was immediately followed by free recall. The paradigm did not involve lexical-decision. The protocol mirrored the FR-LD condition in Chakravarty et al. (2019) experiment 2, except with modified reward contingencies and without the final LD task. Otherwise, the protocol exactly matched the experiment 2 FR-LD condition (see section 4.1).

5.1.1 Value learning task

The learning task simulations followed the same procedure previously described in section 4.1.2 (and see Figure 21) for experiment 2, except with modified reward contingencies. Specifically, on trials showing a 1-point word choosing the "HHHHHH" letter string yielded 40 points (instead of 10). All other reward associations remained unchanged (i.e., avoiding 10-point words still only yielded 1 point). The rewards also remained probabilistic, such that choices yielded the associated point value 90% of the time, and the alternate reward 10% of the time.

5.1.2 Free recall

Free recall simulations followed the same procedure described for Madan and Spetch (2012) (aim 2) described in the section 2.1.6 pseudocode.

5.1.3 Parameter optimization

Model parameters were optimized with the Matlab particle swarm optimizer *particleswarm(),* using 200 particles for all searches. In-house modifications allowed the optimizer to interface with a SLURM scheduler and run particle simulations in a cluster computing environment. The particle swarm optimized $a_U,\ a_W,\ \gamma,\ \lambda,\ \sigma_D,$ and θ_D for the

objective score ${\epsilon_{learn}}^3 + {\epsilon_{recall}}^3$. Importantly, the model was fit to summary statistics from Chakravarty et al. (2019) experiment 2's FR-LD condition. This is important for two reasons. First, the simulation protocol matches the FR-LD condition almost exactly, so learning and memory performance should be highly similar.

Secondly, the FR-LD condition showed no reward-memory differences. Optimizing the model for empirical results without reward-memory effects provides a good test for the hypothesis. The optimizer will attempt to disprove the hypothesis by tuning model parameters with the goal of attenuating reward-memory effects. If the best parameterization (lowest error) still shows reward-memory effects, that indicates the model's predictions are extremely robust to parameter settings (i.e., a strong prediction).

5.2 Results

5.2.1 Value learning

Subjects were simulated with the best parameter values found during parameter optimization (Table 8). Figure 30 shows the simulated subjects' performance in the learning task. Simulated subjects reach performance levels comparable with the experimental data in the FR-LD condition (shown in Figures 23, panel b). The average subject chooses correctly on > 75% of trials by the final block, indicating the model has learned correct behavior for both type of trials (i.e., trials with 1-point and 10-point words).

Table 8. Best parameters found during optimization (see section 5.1.3)

α_U	α_W	γ	λ	σ_D	θ_D
.05	.096	.057	.558	.01	.119

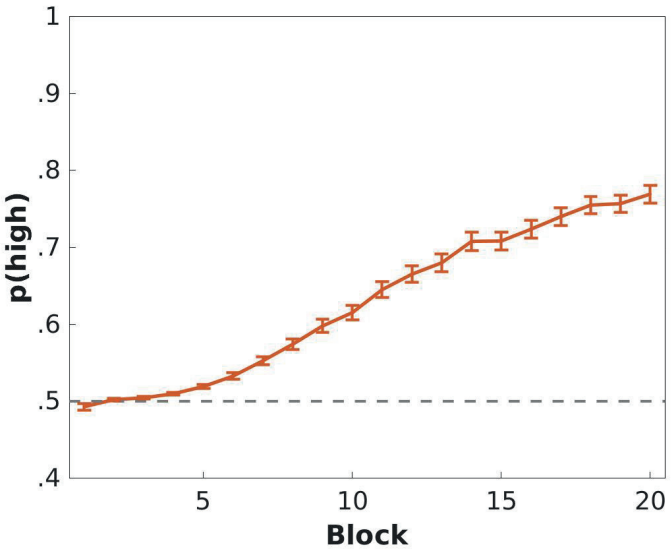

Figure 30. Value learning in novel task simulations. Y axes show probability correct option was chosen.

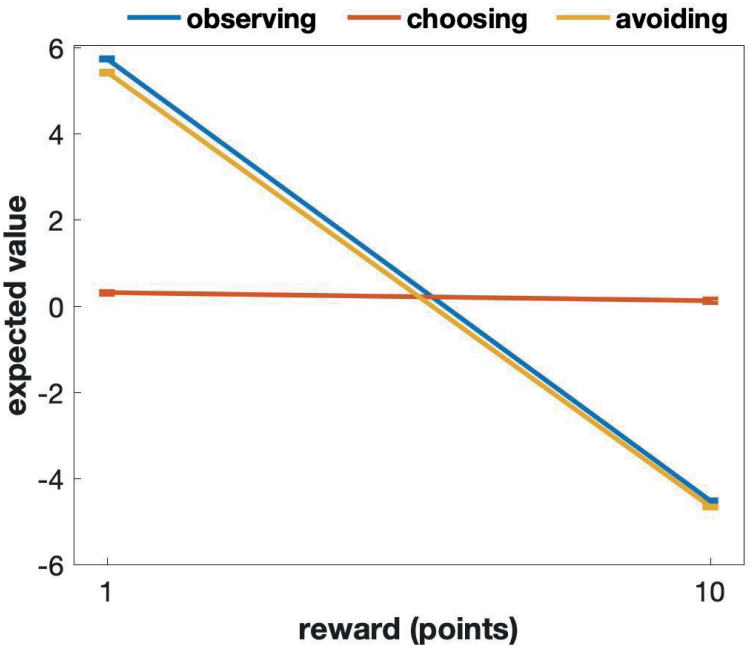

Figure 31. Value function after choice task simulations (novel task). Y axis shows value estimate for features: observing words X(s), choosing words X(a), avoiding words X(a-NOGO).

Figure 31 shows the new reward contingencies unbalanced the value function, as anticipated (compare with figure 27). Subjects expected far greater cumulate reward for features associated with 1-point words than 10-point words. This reflects learning that choosing the alternate "HHHHHH" letter string on trials with 1-point words yields the maximum possible reward (40 points). Furthermore, the negative expected value for avoiding 10-point words reflects learning that choosing 10-point words yields the best reward on those trials (10 points). Figure 31 also shows a negligible difference between expected values for choosing 1-point and 10-point words. The smaller relative difference between receiving 1 and 10 point rewards (e.g., compared to 40 vs. 1) likely explains why the expected value differs so little. Nevertheless, the model clearly learns that trials with 1-point words are far more valuable.

5.2.2 Free recall

Figure 32 shows the simulated free recall performance after value learning, and the model predicts better memory for 1-point words than 10-point words. As hypothesized, the model predicts unbalancing the reward associations will restore the reward-memory effects, even when items values are learned one at a time. The parameters in these simulations were tuned to match the Chakravarty et al. (2019) attenuated memory results, but the model still displayed reward-memory effects. This indicates the model strongly predicts restored memory-effects in this context.

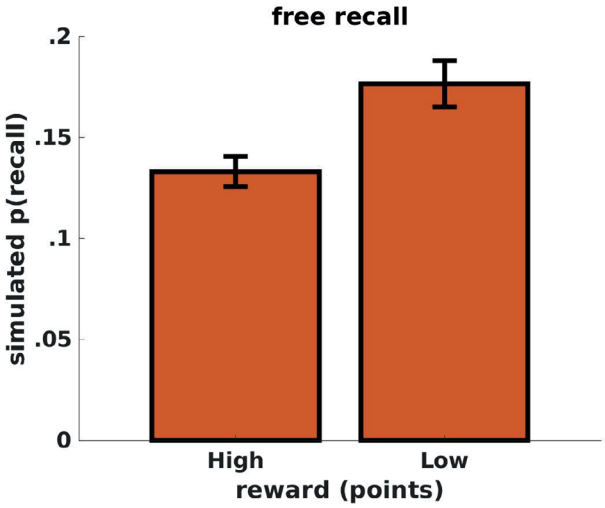

Figure 32. Predicted free recall after novel task. High and low refer to 10-point and 1-point words, respectively.

6. Discussion

The current literature on how reward influences memory shows a complicated pattern of results without a coherent narrative. This modeling effort provides a mechanistic explanation for the data in three reward-memory studies that comprise seven total experiments. I hypothesized that people better remember valuable information, and the modeling results support that account. The models incorporating this hypothesis successfully explained the episodic memory data in all three studies. In contrast, complementary models that did not assume better memory for valuable information failed to capture several important results. This evidence favors the hypothesis that value increases memory strength, and that mechanism underlies the reward-memory effects. This modeling effort also revealed dynamics that crucially support this explanation and provide falsifiable experimental predictions.

The modeling strategy was to start with simple models and step through increasing more complex experiments. The models were incrementally updated with the smallest changes needed for accommodating new experiments. Correspondingly, the initial simulations for experiments in

Madan et al. (2012) (section 2, aim 1) highlighted several fundamental aspects about the hypothetical mechanism in this dissertation. First, the SARSA models illustrate how actions are necessary for explaining reward-memory phenomena. RL models that rely on state value functions rather than state-action values (i.e., Q-learning in SARSA) cannot even learn the 2-AFC task in Madan et al. (2012) without representing the task in artificial ways[4]. In other words, the model requires behavioral information simply to solve the value-learning task in the first place.

Furthermore, these simulations show how integrating behaviors into word memories gives a model where subjects remember valuable information better. This happens because reward-seeking subjects choose the high-reward words more often, and consequently the subject encodes those behaviors more frequently. This produces stronger memories for high-reward words when memory strength also incorporates relevant behaviors (e.g., choosing the word). Figures 1 through 14 demonstrate how this model reproduces the primary memory results in Madan et al. (2012) extremely well, reproducing both the free recall and intrusion error data. However, a complementary model that did not incorporate behavior in memory strength failed to capture these findings. These results support the hypothesis that subjects better remember valuable information, and highlight behavior's importance in this mechanism. A straight-forward prediction from the latter insight is that subjects with chance-level choice performance should not show reward effects in memory; subjects must learn some value function to show differences in memory related to reward.

Madan and Spetch (2012) introduces two new experiments where multiple reward levels now give U-shaped memory outcomes. The U-shaped results are slightly asymmetrical such that

[4] For instance, adding a "state" after each choice that shows the model information about the action

the highest reward words are remembered better than the lowest (figure 17). The simulations in aim 2 (section 3) shows the model also captures these results with two additional terms: NOGO actions and retrieval bias (12-14). Including features for NOGO actions allows the model to explicitly learn values for avoiding behaviors and incorporate that value into relevant word memories. The retrieval bias term simply scales memory strength by the memory's estimated value during recall (i.e., this term increases memory strength for valuable memories). The modeling results showed both terms are necessary to reproduce the asymmetrical U-shaped memory results (figures 17 and 20).

The Madan and Spetch (2012) choice task highlights a counterintuitive situation where words yielding low reward are actually highly valuable. The lowest reward words are never the best choice, so learning to avoid these words confers a valuable strategy (figure 18). This is especially true in situations where the learner may have limited cognitive resources and cannot accurately track all possible stimulus values. The task in Madan and Spetch (2012) is hardly a complex naturalistic decision-making scenario, or even complicated by laboratory task standards. The counterintuitive word values here indicate that many situations (and laboratory tasks) probably also have counterintuitive aspects. Future experiments should consider this perspective when hypothesizing about how reward should influence memory outcomes.

The model required no further changes to accommodate the results in Chakravarty et al. (2019). This study intentionally removed reward-memory effects through a unique choice task described in section 4, so the models only needed to reproduce null effects in these simulated experiments. In hindsight, the data itself offered a less rigorous test in this regard. All models reproduced the data qualitatively, including the models where value was not tied to memory strength. However, Figure 29 shows that removing the retrieval bias term (13-14) degrades the

model's numerical accuracy. The results still show no memory differences between high and low reward words (as observed experimentally), but the precise numerical reproduction suffers. This provides tentative evidence that the Chakravarty et al. (2019) memory results involve some retrieval bias, especially considering retrieval bias was necessary for capturing the U-shaped results in Madan and Spetch (2012). However, that link is certainly tentative.

In Chakravarty et al. (2019) the authors hypothesized that learning word values "in competition" (i.e., via comparison) produces reward-memory effects, and they predicted removing this competition would attenuate reward influences on memory. This idea is not completely at odds with insights gained from the modeling in this study. In fact, the simulations in section 2 demonstrated how systematically choosing high-reward words over low-reward words can explain the memory data in Madan et al. (2012). However, that mechanism alone could not replicate U-shaped memory outcomes; the model also required NOGO behavior and retrieval bias terms. The Chakravarty et al. (2019) simulations also degraded somewhat without a retrieval bias term, and the author's competitive-learning hypothesis should not require retrieval bias.

An interesting aspect of the Chakravarty et al. (2019) choice task is that subjects can earn the maximum possible reward on every trial. Subjects also learn word values one at a time in order to avoid competitive-learning. Consequently, subjects learn that all stimuli and their associated behaviors are equally valuable (Figure 27). This allows models that assume value influences memory to show attenuated reward-memory effects. The words do not differ in value, so the model predicts no memory differences. Altering this aspect of the Chakravarty et al. (2019) choice task provides a test for the mechanism hypothesized in this dissertation, in addition to the competitive-learning hypothesis.

The modeling in section 5 simulated what might happen if trials were not all equally valuable in this task. In these simulations avoiding certain words yielded four times the maximum points possible in Chakravarty et al. (2019), while all other aspects remained the same. Subjects still learned word values one at a time (figure 21), so the competitive-learning hypothesis still predicts no reward-memory differences (i.e., changing the rewards should not change memory outcomes). However, unbalancing the trial rewards should make some words more valuable than others, and this dissertation hypothesizes that value improves memory. The model from aims 2 and 3 (sections 3 & 4) should predict these changes will restore reward-memory effects.

Importantly, these simulations optimized the model parameters for matching the relevant empirical data with no reward memory differences. That is, the simulations attempted to force the model to show no reward-memory effects. This rigorously tested whether the model was so flexible that certain parameterizations could show null results in this situation. Figure 31 shows the optimized model's value estimates after completing the choice task, and the reward changes succeeded in unbalancing the value function. Critically, the simulations also predicted subjects will remember more high-value words (Figure 32)[5].

This result demonstrates a strong experimental prediction given by the current modeling account. The model predicts changing the Chakravarty et al. (2019) reward structure will restore reward-memory effects, and that prediction is robust to different parameterizations. This strong prediction provides excellent falsifiability for the model. Interestingly, the predicted results also clash somewhat with the competitive-learning hypothesis. These simulations maintain the fundamental task structure in Chakravarty et al. (2019); subjects do not learn via comparisons,

[5] Figure 32 x-axis refers to reward, avoiding low-reward words yields the most possible points (highest value)

and behaviors for different words are not mutually exclusive (i.e., choosing one word never comes at the expense of choosing another). However, the model still predicts rewards will influence later memory.

Cumulatively, the modeling in this dissertation shows how subjects' reward-maximizing behavior leads to systematically different actions towards different stimuli. In turn, these different behavioral approaches encode different memories (aim 1, section 2). This can explain how consistently choosing one stimulus over another encodes a stronger memory for the chosen stimulus. However, that cannot account for data in more complicated situations (aim 2, section 3). The model requires two additional considerations. First, avoidance behaviors (i.e., inhibiting an action) can offer the same value as actively taking an action. Second, the value function biases memory strength at retrieval. This retrieval bias gives a distinct mechanism from how reward-seeking behavior influences memory in aim 1 (section 2). Consider a hypothetical scenario where the same behaviors in two situations give different reward outcomes. The model encodes equally strong memories (the behaviors do not differ), but the retrieval bias will make the higher-value experience more memorable during recall.

This idea leads directly to the experimental predictions in section 5. The complex task in Chakravarty et al. (2019) may have attenuated reward-memory effects because the reward structure produced a zero-sum value function (Figure 27), not only because the task eliminated competitive-learning. The modeling in this dissertation suggests both explanations contribute to how reward influences memory, and those contributions depend heavily on the circumstance. Unbalancing the value function in this task should provide further answers on this topic.

6.1 Conclusion

The modeling in this dissertation supports the hypothesis that people better remember valuable experiences. This provides a coherent mechanism that explains how reward influences memory and gives falsifiable predictions. However, the account depends on understanding how value and behavioral strategies differ between situations. Modeling iterations show this quickly becomes difficult as laboratory tasks become even moderately more complex. For example, this work demonstrated how simply adding reward gradations can dramatically change the value function, such that low-reward experiences become extremely valuable. Accommodating these data required the model to assume avoidance behaviors contribute to value estimations, and that value estimates bias memory retrieval. Future work can test the experimental predictions offered here, and further investigate how simpler models can achieve similar dynamics.

7. Appendix

Conceptually, a negative R^2 (i.e., coefficient of determination) value indicates the experimental data's mean provides a better fit than the model predictions. The notation " R^2 " unfortunately suggests the coefficient of determination represents a squared number that cannot be negative, but that is not the case. The definition for R^2 follows

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$

$$SS_{residual} = \sum_i (y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

so R^2 can be negative when the true mean differs from the predicted mean (i.e., $\bar{y} \neq \bar{\hat{y}}$). Researchers who typically use R^2 when evaluating linear models should not encounter negative R^2 often because the intercept term will equate empirical and predicted means (Barten, 1987). However, the model fitting in this dissertation does not ensure $\bar{y} \approx \bar{\hat{y}}$ and therefore we sometimes observe negative R^2. It is important to also note that negative R^2 values do not reflect whether the model captured the data's rank ordering. A model that accurately predicts the relationships between datapoints, but inaccurately predicts the mean, still yields a negative R^2 statistic.

8. References

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron, 50*(3), 507-517.

Barron, H. C., Dolan, R. J., & Behrens, T. E. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat Neurosci, 16*(10), 1492-1498. doi:10.1038/nn.3515

Barten, A. P. (1987). The coefficient of determination for regression without a constant term. In *The Practice of Econometrics* (pp. 181-189): Springer.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nat Commun, 8*, 15958. doi:10.1038/ncomms15958

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nat Neurosci, 20*(7), 997-1003. doi:10.1038/nn.4573

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology, 57*(3), 153-178.

Castel, A. D., Benjamin, A. S., Craik, F. I., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory & Cognition, 30*(7), 1078-1085.

Chakravarty, S., Fujiwara, E., Madan, C. R., Tomlinson, S. E., Ober, I., & Caplan, J. B. (2019). Value bias of verbal memory. *Journal of Memory and Language, 107*, 25-39.

Dabney, W., & Barto, A. G. (2012). *Adaptive step-size for online temporal difference learning.* Paper presented at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci, 8*(12), 1704-1711. doi:10.1038/nn1560

Diederen, K. M., & Schultz, W. (2015). Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology, 114*(3), 1628-1640.

Diederen, K. M., Spencer, T., Vestergaard, M. D., Fletcher, P. C., & Schultz, W. (2016). Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron, 90*(5), 1127-1138.

Duncan, K. D., & Shohamy, D. (2016). Memory states influence value-based decisions. *J Exp Psychol Gen, 145*(11), 1420-1426. doi:10.1037/xge0000231

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *J Neurosci, 38*(33), 7193-7200. doi:10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annu Rev Psychol, 68*, 101-128. doi:10.1146/annurev-psych-122414-033625

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Comput, 24*(6), 1553-1568. doi:10.1162/NECO_a_00282

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585-595. doi:10.1016/j.neuron.2010.04.016

Gluth, S., Sommer, T., Rieskamp, J., & Buchel, C. (2015). Effective Connectivity between Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from Memory. *Neuron, 86*(4), 1078-1090. doi:10.1016/j.neuron.2015.04.023

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General, 136*(3), 389.

Harley, W. F. (1965). The effect of monetary incentive in paired associate learning using a differential method. *Psychonomic Science, 2*(1-12), 377-378.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269-299.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological review, 98*(3), 352.

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature human behaviour, 3*(7), 719-732.

Kensinger, E. A. (2009). Remembering the Details: Effects of Emotion. *Emotion Review, 1*(2), 99-113. doi:10.1177/1754073908100432

Lengyel, M., & Dayan, P. (2008). *Hippocampal contributions to control: the third way.* Paper presented at the Advances in neural information processing systems.

Ludvig, E. A., Madan, C. R., McMillan, N., Xu, Y., & Spetch, M. L. (2018). Living near the edge: How extreme outcomes and their neighbors drive risky choice. *J Exp Psychol Gen, 147*(12), 1905-1918. doi:10.1037/xge0000414

Madan, C. R., Fujiwara, E., Gerson, B. C., & Caplan, J. B. (2012). High reward makes items easier to remember, but harder to bind to a new temporal context. *Front Integr Neurosci, 6*, 61. doi:10.3389/fnint.2012.00061

Madan, C. R., & Spetch, M. L. (2012). Is the enhancement of memory due to reward driven by value or salience? *Acta Psychol (Amst), 139*(2), 343-349. doi:10.1016/j.actpsy.2011.12.010

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour, 1*(9), 680.

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *J Exp Psychol Gen, 145*(5), 548-558. doi:10.1037/xge0000158

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139-154. doi:10.1016/j.jmp.2008.12.005

Norris, D. (2013). Models of visual word recognition. *Trends in cognitive sciences, 17*(10), 517-524.

Preuschoff, K., & Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. *Annals of the New York Academy of Sciences, 1104*(1), 135-146.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory, 2*, 64-99.

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *J Exp Psychol Learn Mem Cogn, 44*(9), 1430-1443. doi:10.1037/xlm0000518

Schacter, D. L., Chiu, C.-Y. P., & Ochsner, K. N. (1993). Implicit memory: A selective review. *Annual review of neuroscience, 16*(1), 159-182.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489. doi:10.1038/nature16961
St-Amand, D., Sheldon, S., & Otto, A. R. (2018). Modulating Episodic Memory Alters Risk Preference during Decision-making. *J Cogn Neurosci, 30*(10), 1433-1441. doi:10.1162/jocn_a_01253
Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning, 3*(1), 9-44.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*: MIT press.
Todd, M. T., Niv, Y., & Cohen, J. D. (2009). *Learning to use working memory in partially observable environments through dopaminergic reinforcement.* Paper presented at the Advances in neural information processing systems.
Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne, 26*(1), 1.
Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., . . . Georgiev, P. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350-354.
Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J Neurosci, 34*(45), 14901-14912. doi:10.1523/JNEUROSCI.0204-14.2014
Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75*(3), 418-424. doi:10.1016/j.neuron.2012.03.042
Zilli, E. A., & Hasselmo, M. E. (2008a). Analyses of Markov decision process structure regarding the possible strategic use of interacting memory systems. *Front Comput Neurosci, 2*, 6. doi:10.3389/neuro.10.006.2008
Zilli, E. A., & Hasselmo, M. E. (2008b). The influence of Markov decision process structure on the possible strategic use of working memory and episodic memory. *PLoS One, 3*(7), e2756. doi:10.1371/journal.pone.0002756