Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations | Graduate Studies

8-2023

Temporally Weighted Averaging: The Effects of Test Delay on Spontaneous Recovery of Choice

Jack Van Allsburg
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

 Part of the Psychology Commons

Recommended Citation

Van Allsburg, Jack, "Temporally Weighted Averaging: The Effects of Test Delay on Spontaneous Recovery of Choice" (2023). *All Graduate Theses and Dissertations*. 8805.
https://digitalcommons.usu.edu/etd/8805

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

TEMPORALLY WEIGHTED AVERAGING: THE EFFECTS OF TEST DELAY ON SPONTANEOUS RECOVERY OF CHOICE

by

Jack Van Allsburg

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

Approved:

Timothy A. Shahan, Ph.D. Thesis Chair	Kerry E. Jordan, Ph.D. Committee Member
Gregory J. Madden, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

Copyright © Jack Van Allsburg 2023

All Rights Reserved

ABSTRACT

Temporally weighted averaging: The effects of test delay

on spontaneous recovery of choice

by

Jack Van Allsburg

Utah State University, 2023

Major Professor: Dr. Timothy A. Shahan
Department: Psychology

Spontaneous recovery of choice is a poorly understood behavioral phenomenon, where, following a delay, animals fail to allocate their behavior in a way that reflects the most recent reinforcement distribution they have experienced, and instead revert to a behavioral allocation consistent with a distribution of reinforcers from the more distant past. This phenomenon, which may play a role in treatment-relevant phenomena such as resurgence, is not predicted by dominant models of dynamic averaging, such as an exponentially weighted moving average. To explore this phenomenon, 3 free-operant experiments with rats were conducted with a serial reversal design in daily 30-minute sessions. The general procedure comprised two phases. In phase 1, two lever responses (A and B) were baited with food pellets on concurrent variable interval schedules, which heavily favored option A (at a ratio of 9:1). During phase 2, lever baiting was reversed to favor option B (at the same ratio of 9:1). At the end of phase 2, rats entered a test delay, where they were maintained at weight in their home cages and no experimental sessions took place. Following this test delay, preference was assessed in test sessions where

levers were presented, but not baited, and the first 2 minutes of responding was used to assess preference between levers. Data analysis included a comparison of dynamic averaging models, including an exponentially weighted moving average, the temporal weighting rule, and several variants of these models. While the data provided strong evidence of spontaneous recovery of choice, the form and extent of recovery was inconsistent with the primary models under investigation. Potential interpretations are discussed, including revised approaches to both the decision rule and valuation functions employed.

(65 pages)

PUBLIC ABSTRACT

Temporally weighted averaging: The effects of test delay on spontaneous recovery of choice

Jack Van Allsburg

Foraging animals in natural environments must track the value of different behavioral options in order to make decisions that maximize their food intake. The process by which they track this value is poorly understood, but holds relevance for our understanding of how animals make choices in general. In a series of experiments conducted in operant chambers, we put rat subjects in a choice scenario where they could press two levers, one of which would intermittently produce the delivery of food pellets on a rich (more frequent) schedule, while the other would do the same on a lean (less frequent) schedule. We switched which lever was richly baited and which lever was leanly baited during a second phase of reinforcement, then imposed test delays where subjects were maintained at weight in their home cages before conducting test sessions in which neither lever was baited. Using this approach, we examined how the relative value of options (as measured by preference between those options) changes over time. Our procedure was explicitly designed to produce a phenomenon called spontaneous recovery of choice, in which a test delay results in a change in preference between options. This phenomenon has been previously linked to relapse in addiction or problem behavior treatment, and is not accounted for by dominant models of how animals track value. Test session data from the three experiments show compelling evidence of this phenomenon, but are inconsistent

with previous work on the topic. Several models are applied to the data and compared quantitatively, and various interpretations are discussed.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

To forage effectively in highly variable natural environments, animals need to make predictions to inform their behavior. High-quality predictions enable animals to make decisions that maximize food intake over time, a maximization theorized to be fundamental to an animal's fitness according to optimal foraging theory (review in Pyke, 1984). However, numerous factors relevant to these predictions exist in a state of constant change—including weather conditions, availability of food, threats of predation or competition, and more. As a result, animals must employ policies for integrating information that is both incomplete and not fully trustworthy (McNamara & Houston, 1980). This high variability also means that as information, such as a memory of prior conditions, will hold less and less predictive value as it ages, because as time passes, conditions are more and more likely to have changed (McNamara & Houston, 1987).

To complicate matters, relevant factors may change at considerably different rates: a given patch might reliably provide nuts or berries for an entire season, but the threat of predation for that patch could vary by the minute or hour—meaning that animals need to track both persistent and transitory trends to make optimal decisions. Models of valuation in foraging, therefore, have historically sought to determine how animals balance the goals of both short-term maximization (reflective of current conditions) and long-term maximization (reflective of historical conditions) (Dow & Lea, 1987).

As a general principle in this compromise, empirical evidence shows animals place more weight on recent experiences, if they have recent experiences to consult

(Cowie, 1977; reviewed in Stephens & Dunlap, 2017), which follows from the notion that recent experience will more accurately describe current conditions and therefore inform optimal decisions. For example, if predicting tomorrow's weather from past experience, a memory of weather conditions from yesterday is more informative than a memory of weather conditions from three weeks ago, because less change will likely have occurred in a single day than in three weeks.

However, the question of how exactly animals track the value of different foraging options—and more specifically, how animals weigh and integrate past experiences to their valuations—remains in debate. The current study aims to evaluate this basis for animals' valuations by comparing the application of two theoretical approaches—an exponentially weighted moving average (Killeen, 1981) and a temporally weighted average (Devenport & Devenport, 1994)—to the behavioral phenomenon of spontaneous recovery of choice.

Exponentially Weighted Moving Average Models

The first approaches to modeling how animals track the value of foraging options over time emerged from the earlier-described compromise between short-term and long-term maximization. An initial theory for how animals incorporate past experience to their decision-making, developed by Cowie (1977), proposes that animals rely on the unweighted average of events within a specific "memory window" of time, and all events older than that window are excluded from this average. The memory window's arbitrary "cutoff" was revised through the development of models based on an exponentially weighted moving average (EWMA, pronounced "yuma") (Killeen, 1981).

In calculating the value of a given option at the current moment (V_n) with this approach (Equation 1), animals assign a set amount of weight, determined by a free parameter β, to their current experience (q_n, the reinforcement received during the current experience), and assign the remainder of weight ($1 - \beta$) to their past experiences (V_{n-1}, their valuation from the last update). Past experiences, therefore, are exponentially discounted in weight each time the model is recursively incremented.

$$V_n = \beta q_n + (1 - \beta) V_{n-1} \quad (1)$$

EWMA models have been adopted widely enough in behavioral ecology, foraging, and decision-making for a EWMA to be termed the "common model" of dynamic averaging (Lea & Dow, 1984). For example, EWMAs have been applied in a wide variety of decision-making or dynamic averaging models by Harley (1981), Myerson and Miezin (1980), Navarro and colleagues (2016), and more.

Beyond explicitly EWMA-based models, however, the implementation of learning rate parameters which integrate past experience with a functionally exponential decay is perhaps even more prevalent, including the family of models based on the linear operator of Bush and Mosteller (1951), such as the Rescorla–Wagner model of classical conditioning (1972). Through the influence of approaches like the Rescorla–Wagner model, exponential decay of past experience has seen widespread adoption in reinforcement learning (reviewed in Katahira, 2015), computational neuroscience (Iigaya et al., 2019; Saito et al., 2014; reviewed in Niv, 2009), and artificial intelligence, in general (Su & Hsu, 2004; Zhang et al, 2013). In short, exponential discounting of past experience or history has become a ubiquitous approach for models of learning across diverse fields and applications.

The Temporal Weighting Rule

As an alternative to EWMA-based approaches that have dominated the field of dynamic averaging, Devenport & Devenport (1993, 1994) introduced the Temporal Weighting Rule (TWR), based on the principle that "time is the source of stochastic change" (Devenport, 1998). TWR puts forth the recency of information as a temporal basis for weighting and entails a different conception of what is required of the brain to compute valuation in making decisions. According to TWR, animals hold a record of their experiences with the alternative (or patch) in memory, including the reinforcement received during each experience and how long ago each occurred. In calculating the value of a given alternative, each experience is assigned weight by how relatively recently it occurred. Equation 2 provides the weight for an experience (W_x) according to TWR.

$$W_x = \frac{1/t_x}{\sum_{k=1}^{n} 1/t_k} \tag{2}$$

The numerator calculates the recency of the experience as the reciprocal of t_x (the time from that experience to the present), while the denominator is the sum of recencies for all n_k experiences with that alternative. Therefore, the weight assigned to each experience is its *relative* recency—the recency of that experience relative to the sum of recencies for all experiences that the animal has had with that alternative. To calculate the value of a given alternative, the reinforcement obtained during each experience is multiplied by that experience's specific weight (as of the current moment), then summed (Equation 3):

$$V = \sum_{x} w_x q_x \tag{3}$$

The valuation (V) equals the sum of reinforcement from each individual experience (q_x) multiplied by the unique weight of that experience (w_x). According to an extension of

Herrnstein's matching law (1961), the relative value of an option theoretically matches the probability of that option being selected (Baum & Rachlin, 1969); e.g. between two options, X and Y, the probability of an animal choosing option X corresponds to its relative value; $P_X = V_X / (V_X + V_Y)$.

A key difference between TWR and other dynamic averaging models is TWR's lack of parameters. While this parameter-free approach affords the model an advantage in parsimony, it provides no means to account for variation in how strongly recency influences the weighting of past experiences—which might be altered by a number of factors, such as individual differences, overall rate of reinforcement (Mazur, 1995), or volatility of reinforcement (Behrens et al., 2007). To address this limitation, Shahan and Craig (2017) introduced a scaled temporal weighting rule (sTWR) (Equation 4):

$$W_x = \frac{1/t_x^c}{\sum_{k=1}^{n} 1/t_k^c} \tag{4}$$

This scalar term changes the degree to which recency influences weighting—or, in more functional terms, changes the relative steepness of the decay function for weights by raising recencies to the power of *c*. The current study incorporates *c* as a free parameter.

Spontaneous Recovery of Choice

Spontaneous recovery of choice is a decision-making phenomenon that represents a challenge to many models of choice behavior and offers a key differentiation in performance between EWMA-based models and TWR. Spontaneous recovery of choice is analogous to the widely studied phenomenon of spontaneous recovery, proper, which has been observed in studies dating back to Pavlov (1927). In Pavlovian conditioning, spontaneous recovery refers to the recurrence of an extinguished association following a temporal delay between the end of extinction and a test of the association. The effect has

been produced by a variety of experimental designs (review in Rescorla, 2004); in a simple example of such a procedure, an initially neutral stimulus (such as a tone or light) is presented to an animal, paired with an unconditioned stimulus (US; such as food) for a period of time. During this initial period of learning, this now-conditioned stimulus (CS) begins to elicit the same response (such as salivation) as the US. This period of time is then followed by a period of extinction conditions, where the CS is presented without the US, and responding gradually decreases until the CS no longer elicits the response. Following this period of extinction, a temporal delay is imposed, during which the animal is not exposed to the CS. After this delay, when the CS is presented again to the animal, despite the association's previous extinction, the animal "spontaneously" resumes responding, recovering the initially learned association between the CS and US.

The fundamental implication of spontaneous recovery is that the initial learning gained during training has not been entirely removed by the period of extinction, and that during the temporal delay, some change is occurring to elicit the recurrence of responding. This characterization of spontaneous recovery can be extended to simple operant conditioning procedures as describing the recovery of a previously extinguished operant response following a temporal delay. As such, spontaneous recovery, whether in Pavlovian or single-response operant procedures, has had various theoretical explanations, including the dissipation of learned inhibition during extinction (Pavlov, 1927), the random replacement of stimulus elements within a conditioned set (Estes, 1955), and differential retrieval of the initial learning resulting from time acting as a contextual cue (Brooks & Bouton, 1993).

Spontaneous recovery of choice, however, differs from spontaneous recovery in that the initial learning is a relative preference for one behavioral option over another option (or options), and the period analogous to "extinction" in classical conditioning or a single-response procedure may simply be a period of different conditions—not necessarily a lack of reinforcement. The temporal delay, in this circumstance, then elicits the recovery of an initially learned preference between options, rather than recovery of extinguished responding.

Two studies that illustrate spontaneous recovery of choice were performed by Mazur (1995, 1996). The first study (Mazur, 1995) comprised a series of experiments in which pigeons were trained to peck two keys, which were initially reinforced on equal concurrent variable interval (VI) schedules. After several daily 30-min sessions with equal schedules, one of the keys was assigned a richer schedule, and over the course of the sessions following this switch, subjects adjusted their behavioral allocation so that their response proportions asymptotically approached the new reinforcement proportions. However, at the beginning of these post-switch sessions, subjects' response proportions initially reflected the reinforcement proportions of pre-switch sessions, rather than the proportions of the most recent session. In other words, following the 23.5-hr inter-session interval, subjects partially recovered the preference they acquired during earlier sessions, rather than maintaining their preference from the end of the previous session.

A further investigation (Mazur, 1996) followed a similar procedure, except switching one key to the richer schedule for only one, two, or three sessions before reverting to equal proportions of reinforcement. Again, subjects' preferences at the beginning of sessions appeared to reflect earlier sessions, both during the switch to a

richer schedule and the switch back to equal reinforcement proportions. Additionally, the number of richer schedule sessions experienced was positively correlated with the strength of influence those sessions appeared to have on initial preferences in sessions after the reversion to equal reinforcement proportions. In one experiment, the introduction of a 3-day "rest period" was shown to attenuate the influence of past sessions when inserted between equal reinforcement proportions sessions and the sessions with a richer schedule for one key. This latter finding suggests a specific relationship between recency and strength of influence for past experiences.

However, in comparison to the literature on spontaneous recovery of a single response or association, the basic empirical properties of spontaneous recovery of choice have been much less specifically studied. In particular, the relationship between test delay and extent of recovery has been widely studied for simple spontaneous recovery, but few studies have been conducted on this relationship for spontaneous recovery of choice. For spontaneous recovery of a single response or association, the extent of recovery associated with increasing test delay has been shown to consistently follow a positive, negatively accelerated curve in studies using a variety of species and a variety of procedures. One of the first studies to document this relationship found that when rats were trained on a single bar-press response that was later extinguished, groups assigned longer test delays showed greater recovery, but that the overall trend in recovery was negatively accelerated as test delay increased (Ellson, 1938). A similar trend in recovery was found for a classically conditioned eyeblink response with humans (Grant et al., 1958), a classically conditioned eyeblink response with rabbits (Haberlandt et al., 1978),

a classically conditioned fear response with rats (Quirk, 2002), and a single-response operant procedure using pigeons (Robbins, 1990).

Understanding the relationship between test delay and spontaneous recovery of choice is highly relevant to our understanding of how animals dynamically value different options for two primary reasons. First, if the passage of time alone can produce a significant change in preference between the two options, it would suggest, at the least, that there is a fundamentally temporal dimension to the valuations or decision rules employed by animals in allocating behavior. Second, if the passage of time produces an increasing degree of recovery (like the negatively accelerated curves observed in single-response and Pavlovian procedures), the specific form of this function may provide insight into potential mechanisms of valuation, such as a temporally weighted average, that could underlie such a change in behavior.

Applying TWR and EWMA Models to Spontaneous Recovery of Choice

TWR and EWMA models crucially differ in how the weights assigned to past experiences change over time, and one of the most effective illustrations of this contrast is found in spontaneous recovery of choice. According to TWR, because recency is calculated as $1/t_x$, the passage of time (increasing time, t, since the experience) means recency will decay hyperbolically. As a result, if time continues to pass without new experiences, *relative* differences in recency between experiences will become increasingly smaller, and valuation will approach the unweighted average of reinforcement received during those experiences (Figure 1). For example, as depicted in Figure 1, say that animals experience three sessions where reinforcement favors option A (at a ratio of 9:1, relative to option B), followed by two sessions favoring option B (at a

ratio of 9:1, relative to option A), and no new experiences with the alternatives occur following these sessions. On the day after the final session (day 6), if an animal is weighing the fourth session (2 days before, with a recency of 1/2) vs. the fifth session (1 day before, with a recency of 1/1), there is a substantial relative difference between them (1/2 vs. 1, ratio of 1:2). However, if we consider those same experiences 15 days after the final session (day 20); the fourth session now has a recency of 1/16 and the fifth session now has a recency of 1/15—shrinking the relative difference between them (1/16 vs. 1/15, ratio of ~1:1.07). By contrast, the weights functionally assigned by a EWMA model will decay at the same exponential rate, preserving the relative differences in valuation between options at days 6, 10, and 20, regardless of whether the model is updated based on new experience or the passage of time (Figure 1).

The qualitative explanation of spontaneous recovery of choice by TWR is therefore relatively straightforward. If a period of reinforcement (like sessions 1-3 in Figures 1 and 2) favoring behavioral option A is followed by a shorter period favoring behavioral option B (like sessions 4-5 in Figures 1 and 2), valuation just after the latter period will assign more weight to the experiences favoring option B—leading to a comparatively higher valuation of option B, and a low preference for option A (P_A; see day 6 in Figure 2).

However, as time passes, and the relative difference in recency between the two periods becomes less significant, valuations will approach the unweighted averages of reinforcement from each option. Accordingly, the relative weight assigned to experiences favoring option B will decrease and the relative weight assigned to experiences favoring option A will increase (Figure 1)—leading to a higher comparative valuation of option A

and a corresponding recovery of preference for option A (see day 20 in Figure 2). By contrast, a EWMA model's weightings will remain static (Figure 1), and will not predict spontaneous recovery of choice, regardless of test delay (see days 6 – 20 in Figure 2).

Previous studies on Spontaneous Recovery of Choice and TWR

Because spontaneous recovery of choice is an effect for which TWR offers a distinct and mechanistic explanation, many of the studies specifically related to this phenomenon were performed as tests of predictions of TWR. Because models based on a EWMA do not predict this effect, each of the studies reviewed inherently contradicts predictions from those models.

As an initial evaluation of TWR's application to choice behavior, Devenport and Devenport (1994) performed a study using wild golden-mantled squirrels and least chipmunks designed to produce spontaneous recovery of choice. During an initial training period, animals were able to retrieve food from two feeding stands, one of which was baited with unhulled sunflower seeds. For a "stable" group (about half the subjects), training concluded with this initial period. For a "variable" group (the remaining subjects), this initial training was followed 2 hr later by a period where the baiting conditions were reversed (the originally baited stand was empty and the originally empty stand was baited). Roughly the same number of visits were allowed during the two training phases. Stand preference was assessed for both groups at an early (1 hr after end of training) and late (48 hr after end of training) test time by recording the subjects' first choice between the two stands. Predictions for these tests were calculated using an unweighted average of a hypothetical animal's experiences with the training phases and an average weighted according to TWR as of the two test time points. These weighted

averages used the time between the test and the midpoint of the training phase(s) to calculate the weights of those experiences and their reinforcement. The “stable” group showed consistent and exclusive preference for the initially baited stand. The “variable” group, by contrast, showed a shift in preference, where the more recently baited stand was exclusively preferred at the early test (proportion of first choices to the more recently baited stand = 1.0), and preference was roughly even between the two stands at the later test (proportion of first choices to the more recently baited stand = .54). Because this group showed exclusive preference (1.0) for the more recently baited stand at the early test, these results were inconsistent with predictions from an unweighted average or chance (both of which would predict .5). Because this group then showed indifference between the two stands at the late test (.54), these results are also inconsistent with choices based on the most recently baited stand (which would predict 1.0). A second experiment found similar results on a longer timescale using otherwise the same procedure. In a third experiment, conditions of reinforcement were varied more frequently, such that over the course of a five-day training phase, at each hour the location of food might change, and one of the two experimental stands was baited for twice as many hours as the other. Following a delay, chipmunks maintained their preferences for the stand that had previously been richer on average rather than the stand that had been more recently baited, suggesting that the delay alone has not produced a reversion to an exploratory mode of foraging.

Devenport and Devenport (1993) obtained similar results with a similar procedure using dogs—subjects chose based on more recent information when available, but recovered their preference consistently with a regression to unweighted averages

following the imposition of a temporal delay. Similar results were also obtained in studies with horses (Devenport et al., 2005), suggesting that these valuation mechanisms may also be employed by grazing herbivores. In several studies examining spontaneous recovery of spatial preference, longer test delays were found to elicit spontaneous recovery in rats (Devenport, 1998) mice (Lattal et al., 2003) and pigeons (Leising et al., 2015). In the previously summarized studies of Mazur (1995, 1996), the recovery of preferences following inter-session intervals was also consistent with a temporally weighted average like TWR. Additionally, the introduction of a three-day "rest period" in the latter study (1996) also produced changes qualitatively consistent with a temporally weighted average. However, the extent of recovery found by Mazur did not follow the quantitative predictions of TWR, but a scaled version of the model was not tested, so further investigation will be needed to evaluate sTWR's account of the data.

In a later study using rats, Devenport and colleagues (1997) performed a systematic study of test delay and spontaneous recovery of choice. In a trial-based procedure, subjects were able to choose between two patches, located in opposite corners on the far side of a rectangular platform. During training, a removable barrier was placed to bisect the platform such that subjects could only access one patch at a time, and subjects were provided eight opportunities to alternatingly access each patch (16 trials total). During these first 16 trials, patch A (assignment randomly counterbalanced) was baited, while patch B was not baited. During an ensuing 16 trials, baiting conditions were reversed so that patch B was baited and patch A was empty. While both groups were provided access to an equal number of total pellets during these 32 trials, groups in the experimental condition A = B found the two patches baited with equal numbers of pellets

(ratio of 1:1), while groups in the other experimental condition A > B found patch A baited with far more pellets than patch B (ratio of 5:1). A series of four different delays (1, 240, 360, and 1440 min) were imposed before subjects were given the opportunity to choose between the two patches, both unbaited. The 1-min delay groups for both A = B and A > B conditions exclusively chose patch B, which had been most recently baited. As delays increased however, groups in both conditions regressed to choices consistent with the unweighted average values for the patches, with the A = B groups showing indifference at the two longest test delays and A > B groups showing exclusive preference for patch A at the two longest test delays.

The Current Study

The current study comprises a series of three experiments, thoroughly examining the relationship between test delay and spontaneous recovery of choice. As described above, the most directly comparable study of this relationship by Devenport and colleagues (1997) found recovery consistent with studies of test delay and spontaneous recovery in classical or single-response operant procedures. However, this finding was demonstrated only on a short timescale, with 24 hr as the longest delay, and with a discrete, forced-choice training procedure. By contrast, the current study examines this question in a free-operant procedure on a much longer timescale. These differences allow for more apt application to human choice paradigms, where changes in behavior over days and weeks may be more relevant to effective treatments than changes over a single day. Also, because alternatives were always available, rather than presented exclusively (as in the forced-choice procedure) the results are more readily translated to realistic choice scenarios.

Experiment 1 constitutes a translation of the original Devenport and Devenport studies (1993,1994) to a free operant laboratory procedure, using a within-subjects design. Experiment 2 revises some procedural parameters of Experiment 1 in an attempt to determine if greater spontaneous recovery of choice might be produced. Experiment 3 replicates key aspects of Experiment 2, but employs a between-subjects design to resolve potential confounds due to the within-subjects design of Experiments 1 and 2.

CHAPTER II

GENERAL METHODS AND PROCEDURE

Subjects

Experimentally naive male Long-Evans rats (approximately 72-92 days old) served as subjects. Rats were individually housed in a colony room controlled for humidity and temperature and illuminated on a 12:12 hr light/dark cycle. Subjects were maintained at 80% their free-feeding weight and provided access to water (ab libitum). Sessions were conducted at approximately the same time each day, 7 days/week.

Apparatus

Ten identical operant chambers (30 cm x 24 cm x 21 cm; Med Associates) were housed in sound- and light-attenuating cubicles. Chambers included work panels on the front and back walls, with a clear Plexiglas ceiling, door, and wall opposite the door. On the back wall, a centered house light provided chamber illumination. On the front wall, two retractable levers were positioned on either side of a food magazine. Med-PC software controlled all experimental events and data collection.

Procedure

Magazine and Lever Training

Experiments began with one session to train subjects to retrieve pellets delivered by a magazine. 30 pellets were delivered via magazine on a 60 s variable-time schedule, which illuminated for 3 s with each delivery. In four subsequent sessions, rats were trained on the lever press response. During these lever training sessions, the house light illuminated to signal the session's beginning. Simultaneously, one lever extended and

each lever press caused the lever to retract, the house light to extinguish, and a food pellet to be delivered. The delivery of a pellet initiated a 3 s consumption period, during which the magazine was illuminated, and the lever retracted. After this period, the same lever extended, and the house light was illuminated. For half of the subjects, the lever for option A extended during the first and third lever training sessions. For the remaining subjects, the lever for option B extended during the first and third sessions. The second and fourth sessions were used to train subjects on the opposite lever using the same method. Option A and B lever assignment was counterbalanced across subjects. Sessions terminated once 100 pellets were earned.

Phase 1 Sessions

In Phase 1, daily sessions began with the house light illuminating and both levers extending. In the first session, the first lever press on either lever resulted in a pellet delivery. Following this first press, and for all other sessions during this phase, levers were baited on concurrent VI schedules. The option A lever was baited on a VI-10 s schedule while the option B lever was baited on a VI-90 s schedule. All VI schedules were constructed using ten intervals derived from the Fleshler and Hoffman (1962) distribution. To prevent animals from employing a simple alternating strategy, a 3 s changeover delay (as in Baum, 1982) was employed (following each response, responses on the opposite lever did not produce pellet deliveries for a period of 3 s). Also, the delivery of a pellet initiated a 3 s consumption period, during which both levers retracted, and the session timer paused. Following this period, the levers extended again, and the timer resumed. Sessions terminated after 30 min. The number of daily sessions in Phase 1 varied between the three experiments (detailed in each experimental chapter).

Phase 2 Sessions

Beginning the day after Phase 1 ended, Phase 2 daily sessions continued in the same manner as Phase 1, but with reversed baiting conditions. That is, the option A lever was baited on a VI-90 s schedule while the option B lever was baited on a VI-10 s schedule. The changeover delay and 30-min duration remained the same. The number of daily sessions in Phase 2 also varied between the three experiments (detailed in each experimental chapter).

Testing Phase Sessions

The testing phase comprised 1 or 2 test sessions for each group following various test delays (specific delays and tests detailed in each experimental chapter). During test delays, animals remained in their home cage and maintained at their current weight. During test sessions, the house light illuminated to signal the beginning of the session, and both levers extended, but neither response was baited. The first 2 min of responding in each test session was used in data analysis to assess initial preference (responses on option A divided by total responses).

CHAPTER III

EXPERIMENT 1

This experiment examined the effects of test delay on spontaneous recovery of choice by testing subjects at two different test delays to observe changes in relative preference following two phases of different reinforcement conditions. This within-subjects design followed the approach of previous studies of spontaneous recovery of choice (Devenport & Devenport, 1993; Devenport & Devenport, 1994). The procedure exposed subjects to reinforcement conditions favoring option A during a longer Phase 1 before reversing reinforcement conditions to favor option B during a shorter Phase 2. By testing subsequent preference at two different test delays, the design aimed to isolate the effect of time's passage (i.e., test delay) on subjects' relative valuation of options. The occurrence of a significant increase in the relative valuation of option A served as our criterion for evidence of spontaneous recovery of choice behavior.

Method

10 rat subjects underwent magazine and lever training before the three phase, free-operant serial reversal procedure detailed above (see Chapter II: General Methods and Procedure). The length of Phase 1 (in which reinforcement proportions favored option A at a ratio of 9:1) and Phase 2 (in which reinforcement proportions favored option B at a ratio of 9:1) in this experiment was determined by simulating predictions for preference using the unscaled TWR model. Various lengths (ranging from 2 to 30 days) for both phases were simulated and a combination of phase lengths that maximized both the predicted reversal (during the phase of reversed conditions) and subsequent

recovery of preference (following the test delay) was selected (Phase 1: 18 days; Phase 2: 7 days). Two tests were performed at two different test delays. Test 1 occurred on day 26, immediately after the final session of Phase 2 (a 1-day delay), and Test 2 occurred on day 50, after a delay of 25 days (equal to the summed lengths of Phases 1 and 2).

Results

Subjects acquired a preference for option A (P_A) of .940 by the end of the first phase of 18 daily sessions, then reversed this preference (P_A = .131) by the end of the second phase of 7 daily sessions (see full results in Figure 3). At Test 1, conducted on day 26, a low preference for option A was observed (P_A= .199). At Test 2, conducted on day 50, preference had increased to P_A = .372, but it did not rise above indifference between the two options (.5). Because preference data was calculated as proportions, and therefore bounded between 0 and 1, the data were transformed to logits (log odds) to avoid violating assumptions of parametric statistical analysis (an overview of logits can be found in Cramer, 2003). Logits are calculated simply by taking the natural log of the ratio of a proportion (p) to its complement (1-p). Proportions ranging from 0 to 1 will therefore range from negative to positive infinity as logits, better approximating a normal distribution of the data. The difference between preference in logits was statistically significant ($t(9) = 5.81$, $p < .001$). Three models were fitted to preference data throughout the experiment. Model predictions from TWR, sTWR, and a EWMA model were calculated using the programmed rates of reinforcement. The Solver extension in Microsoft Excel was used to fit sTWR and a EWMA model to the data, adjusting the parameters c and β, respectively, in minimizing residual sum of squares (employing the GRG-Non-linear algorithm). Observed data were better described by predictions from a

fitted sTWR (c = 2.27, R^2 = .982) than from an unscaled TWR (R^2 = .128) or from a fitted EWMA model (β = .603, R^2 = .964). It is important to note that while the EWMA model provided a better fit overall than TWR, it failed to predict the observed recovery. Information criterion comparison of these models favored sTWR (*AIC* = -102.45, *BIC* = -101.56) over the EWMA model (*AIC* = -90.13, *BIC* = -89.24) with an evidence ratio of 8428:1 (see Klapes et al., 2018).

Discussion

Behavior during Phases 1 and 2 approximated matching, with behavioral proportions roughly equal to reinforcement proportions by the end of each phase. In Phase 1, subjects acquired a strong preference for the richer alternative, option A, before gradually reversing this preference during Phase 2, where option B was the richer alternative. All three models (EWMA, TWR, and sTWR) accounted for this behavior well, as expected, but their performance differed for the two test sessions. The significant difference between the observed preference at Tests 1 and 2 provides clear and compelling evidence of spontaneous recovery of choice following the passage of time.

On its face, the occurrence of spontaneous recovery of choice acts as evidence against the account provided by a EWMA model of valuation, and potentially in support of an account by a weighted average like TWR or sTWR. However, the result fell far below the predictions of TWR for this later test and was much better described by sTWR (with a fitted parameter value of c = 2.27). The poor performance of TWR in describing these data suggest that the parameter c is indeed necessary to account for differences in scaling (such as in timescale). Given sTWR's relative success in describing the data, we hypothesized that revising the lengths of the two initial phases to correspond with the

predictions of sTWR could produce more significant recovery of choice and incorporated this insight to the design of Experiment 2.

CHAPTER IV

EXPERIMENT 2

Experiment 2 followed the same basic design and procedure of Experiment 1, only changing the lengths of Phase 1 and Phase 2 based on the results of Experiment 1. All other aspects of the procedure were replicated, including within-subjects comparison.

Method

10 rat subjects again underwent magazine and lever training before the above-described three phase, free-operant serial reversal procedure. To determine phase lengths for Experiment 2, predictions for preference from sTWR (with the parameter value that best fit the data from Experiment 1: $c = 2.27$) based on various lengths (ranging from 2 to 30 days) for both phases were simulated, and phase lengths (Phase 1: 14 days; Phase 2: 2 days) were selected by the same criteria as Experiment 1: maximizing the predicted reversal and subsequent recovery of preference.

Results

Subjects again acquired a strong preference for option A ($P_A = .917$) by the end of the first phase, then reversed this preference for option A ($P_A = .150$) by the end of the second phase (see full results in Figure 4). Test 1 was conducted the day after the second phase, and a low preference for option A was observed ($P_A = .291$). Test 2 was conducted after a test delay of 16 days from the end of the second phase, and although preference significantly increased to .534 (t-test of logit preference between tests: $t(9) = 5.162$, $p < .001$) from Test 1, it remained within a standard error of the mean (.043) of indifference (.5). Model predictions from TWR, sTWR, and a EWMA model were calculated using the programmed rates of reinforcement and fit by the method described in the results

section of Chapter III. The observed data were again better described by a fitted sTWR (c = 2.53, R^2 = .947) than TWR (R^2 = .044) or a fitted EWMA (β = .514, R^2 = .895). In contrasting sTWR and EWMA, the information criterion comparison favored sTWR (*AIC* = -165.03, *BIC* = -163.73) over EWMA (*AIC* =-146.95, *BIC* = -145.66; evidence ratio of 473:1).

Discussion

All three models again accounted for well for Phases 1 and 2, but performance differed for the two test sessions. The significant difference between the observed preference at Tests 1 and 2 provided further evidence of spontaneous recovery of choice following the passage of time, and further support for the account of sTWR over the account by a EWMA model of valuation.

However, while the observed Test 1 preferences in both Experiments 1 and 2 fell within a standard error of the predicted preferences, observed preferences at Test 2 were considerably lower than predictions from sTWR or TWR. This discrepancy, appearing in both experiments, suggests the possibility of a testing effect, with the experience of Test 1 potentially influencing performance during Test 2. During testing, subjects experienced a 30-min session during which previously reinforced alternatives are no longer reinforced. So, while the relative preference between the valuations of the two options may hypothetically have been preserved, this period of extinction could be affecting other factors relevant to performance of responses at the later test.

The notion that extinction conditions during a test might influence later testing is not novel. Skinner (1938), in a single operant response procedure, performed two tests under extinction conditions for the recovery of an extinguished response, the first test

occurring the day after the response had been extinguished, the second test occurring 43 days later. For a group that experienced both tests, spontaneous recovery was significantly attenuated at the later test, relative to a group that was only tested at the later time. This difference (between the group that tested twice and the group only tested later) suggests that performance during the later test may have been influenced by the extinction conditions experienced during the first test. In a similar single-response procedure, Ellson (1938) tested recovery of a single bar-pressing response after the response had been extinguished for four groups, each tested with a different delay: 5.5, 25, 65, or 185 min. The groups' responding during those tests fell along a negatively accelerated curve, consistent with the predictions of valuation from a temporally weighted average like TWR. In short, we hypothesized that investigating recovery with a between-subjects design may isolate the effect of time delay from the effect of continued extinction, a hypothesis we then tested with the between-subjects design of Experiment 3.

CHAPTER V

EXPERIMENT 3

Experiment 3 examined the relation between time delay and spontaneous recovery of choice using a between-subjects comparison of preference, exposing all subjects to the same history of reinforcement, but instituting a different test delay length for each of four groups. Because a testing effect could potentially have moderated the relationship between recovery and test delay in the within-subjects designs of Experiments 1 and 2, a between-subjects design was employed to provide a more robust test of the quantitative predictions of sTWR and a direct comparison of sTWR and a EWMA model for the relationship between test delay and recovery of choice.

Method

40 rat subjects underwent magazine and lever training before the above-described three phase, free-operant serial reversal procedure. Phase 1 and 2 lengths were equivalent to those used in Experiment 2 (Phase 1: 14 days; Phase 2: 2 days). At the end of Phase 2, rats were assigned to 4 groups such that average preference during the last two days of Phases 1 and 2 did not significantly differ between groups (Phase 1 logit terminal preference comparison ANOVA: $F(3,35) = .156$, $p = .931$; Phase 2 logit terminal preference comparison ANOVA: $F(3,35) = .089$, $p = .966$). Group 1 was tested with a 1-day test delay (testing on day 17, the day after phase 2), Group 2 with a 3-day test delay (testing on day 19), Group 3 with an 8-day test delay (testing on day 24), and Group 4 with a 32-day test delay (testing on day 48).

Results

Regression Analysis

Observed preference data during the testing phase (see Figure 5), showed a monotonic increase as a function of test delay. However, the data failed to take the negatively accelerated curvilinear form found in previous studies of spontaneous recovery of choice (such as Devenport et al., 1997). While this discrepancy will be discussed and interpreted in greater detail below, visual analysis on the limited number of tests failed to suggest a specifically curvilinear form for the relationship. Given this uncertainty in form, a linear regression analysis was conducted to quantitatively assess the significance of the relationship between test delay and preference.

A linear regression fit of the data revealed a significant, positive relationship between delay and preference in logits ($p = 0.004$, $R^2 = .20$). Test session data as proportions and logits can be found in Table 1, while the summary statistics of this linear regression fit can be found in Table 2. Because the animals received no new experience during the testing phase until their test session, the role of delay during this time as a significant positive predictor of preference constitutes strong evidence for spontaneous recovery of choice. However, it is notable that recovery never increased above indifference (indifference—a proportion of .5, transforms to a logit value of 0), similarly to the data gathered in Experiments 1 and 2.

Model Comparison

Model predictions from TWR, sTWR, and a EWMA model were again calculated using the programmed rates of reinforcement to allow for unified predictions across all groups and fit by the method described in the results section of Chapter III. Unlike in Experiments 1 and 2, sTWR failed to outperform the EWMA model in accounting for the

data obtained. The EWMA model (β = .476, R^2 = .940) showed a better fit than sTWR (c = 3.15, R^2 = .876) or TWR (R^2 = .158). Similarly, information criterion comparison favored the EWMA model (AIC = -108.71, BIC = -107.71) over sTWR (AIC = -94.09, BIC = -93.09), with a ΔIC of 14.61, constituting an evidence ratio of ~1488:1 in favor of the EWMA model.

Discussion

Again, all three models under primary investigation described behavior during Phase 1 and Phase 2 well. Behavior during the test sessions, by contrast, was not well modeled by any of the three models. Preference during these sessions remained relatively constant for the first three tests, then increased to roughly approximate indifferent responding. Although our model comparison favored the EWMA model, this result is somewhat misleading. While it is true that the data are better fit by the EWMA model than TWR or sTWR over the course of the experiment, the test session data stand in stark disagreement with the EWMA model's account of the effect of test delay. Since EWMA models of valuation do not predict changes in preference without new experience, the observed evidence of statistically significant spontaneous recovery of choice is fundamentally inconsistent with a EWMA-based account. Simultaneously, while the observed recovery is qualitatively consistent with an account from a weighted average like TWR or sTWR, these results are quantitatively inconsistent with the form and extent of spontaneous recovery of choice predicted by TWR or sTWR—to the point that the static preference predicted by the EWMA model is better supported by the comparison. In short, none of the three primary models under investigation were decisively more effective in describing the test session data.

Taking a step back, the conundrum of interpreting Experiment 3's results might be best understood by examining three salient features of the test data in turn: 1) preference increased at the shortest test delay of 1 day; 2) preference did not increase for the intermediate test delays of 3 and 8 days; and 3) preference increased again at the longest test delay of 32 days. Since the account from TWR could be provided by sTWR (if the best fit value of c was 1), and this account was roundly outperformed in the model comparison, this examination will focus on contrasting the accounts of a EWMA model and sTWR in explaining these three features—for clarity and simplicity.

The first feature, the increase from the end of Phase 2 to the first test (1 day delay), poses challenges for EWMA models, because preference at this first test falls well above what would be predicted by a EWMA model fit to just the data from Phase 1 and II: the model (with a fitted β value of .476) predicts a preference of .199, while at the first test, subjects produced an actual preference of .339 (SEM = .047.). If one were to perform the same calculation using sTWR, the model (with a fitted c value of 2.19) predicts a preference of .372. Therefore, this first feature of the data, at least on its face, seems to better support an sTWR account.

On the other hand, the second feature of the data, the lack of increase at test delays of 3 and 8 days, is inconsistent with sTWR. Subsequent test session predictions from sTWR (fit to Phase 1 and 2 alone, $c = 2.19$) increase above indifference (.5), while the observed preference data show no such increase. Taken in isolation, this lack of elevation is qualitatively consistent with a EWMA account, but already, neither model provides a compelling quantitative account of the data.

The third feature of this data, the increase seen at the final test (32 days), is difficult to explain by either model. From the perspective of a EWMA account, this sort of increase directly contradicts a static account of preference, even if we ignored the increase from the end of Phase 2 to the first test. From the perspective of an sTWR account, this late increase is also unexpected, given that sTWR would predict a negatively accelerated curve of recovery. We conclude that neither of these models in their current formulation describe the test data particularly well.

In addition, we found little evidence to suggest that the potential testing effect which shaped the design of Experiment 3 actually occurred. Preference again failed to rise above indifference, as in Experiment 2, even at the latest test. Even more, we conducted an additional test with all of the first three test groups on day 48 to compare their preference during a second test to the preference of the group only tested once and found no significant difference between groups with a one-way ANOVA of logit preference ($F(3, 35) = 0.105$, $p = 0.957$).

CHAPTER VI

GENERAL DISCUSSION

Together, these three experiments outline a consistent, but difficult-to-interpret finding: spontaneous recovery of choice can be reliably produced by time's passage using the serial reversal procedure and test delays employed—but the form and extent of that recovery is poorly described by our current approach to modeling it. Further, these results are inconsistent with the most comparable previous study of the phenomenon (Devenport et al. 1997). In Experiments 1 and 2, exposing subjects to a longer test delay produced a significant increase in preference. While these experiments may have faced some confounds due to their within-subjects design, Experiment 3 similarly found a significant positive relationship between test delay and preference with a between-subjects design. However, at no point in any of the three experiments did recovery significantly rise above indifferent responding (equal allocation between options), and the recovery obtained was poorly described by the models under investigation.

While our models' failure in description may result from the inaccuracy of these approaches in describing the valuation of options—it could also result from an incorrect formulation of the decision rule employed for choosing between options. In essence, we must now ask if the failure of these models should be attributed to an inadequate account of valuation (modeling the process of integrating experience) or an inadequate account of choice (modeling the decision kernel guiding allocation based on calculated valuations).

Valuation functions

Let us first turn to potential interpretations of these data that relate to the valuation functions employed. In all three experiments, we tested the performance of valuation models by evoking spontaneous recovery of choice. This effect, if we accept our current decision rule as accurate (for the moment), would provide evidence of a specific characteristic of how these functions calculate value: as experience fades into the past, its weight in decision-making decays at a decreasing rate. The weights assigned to past experience by TWR or sTWR produce a hyperbolic decay of this form: experience loses weight more and more slowly as time goes on. By contrast, EWMA models functionally assign weight that decays at a constant exponential rate—meaning that relative preference should remain static without new experience. The fact that spontaneous recovery of choice occurred in all three experiments serves as evidence against a EWMA account. At the same time, in Experiment 3, which provided higher resolution of this recovery, TWR and sTWR failed to predict the form of the data, and were actually outperformed by the EWMA model, suggesting neither of the approaches under primary investigation were effective in modeling these data.

However, there is a class of valuation models that does relatively well in accounting for these data, with some caveats. Recent work in neuroscience has promoted the use of a multiple-timescale version of the EWMA model (Iigaya et al., 2019). The simplest version of such a model actually employs two EWMAs, a "fast" integrator with a larger β parameter (e.g. β_x; corresponding to a shorter timescale), and a "slow" integrator with a smaller β parameter (e.g. β_y; corresponding to a longer timescale). The model then divides weight between the two integrators with a third parameter (W_x). In effect, this arrangement allows the weight of past experience to decay at a variable rate,

approximating the same sort of declining decay rate found in the hyperbolic weightings of sTWR or TWR.

As previously discussed, this declining decay rate in the weightings of past experience makes a 2-timescale EWMA model capable of describing preference reversals and spontaneous recovery of choice. Indeed, when a 2-timescale EWMA model is fit to data from Experiment 3, we find the best performance so far (R^2 = .979; see Table 3 for full model comparison).

Description can be improved even further with a 3-timescale EWMA model. This model, by employing three integrators, requires 3 learning rate parameters (e.g. β_x, β_y, β_z) and 2 parameters to determine the weight assigned to each of the 3 integrators (W_x, W_y). The fit produced by this 3-timescale EWMA model closely describes the data (R^2 = .988).

However, the fits from these many-parameter models are somewhat misleading. Behavior in Phase 1 and 2 of Experiment 3 approximates simple matching of behavior to reinforcement proportions[1], meaning that testing acts as the crucial period of performance for these models. With only 4 test sessions, models with more parameters are simply more capable of closely describing the data in post-hoc analysis. For example, we could take the same approach to sTWR as the multi-integrator EWMA models, and use two valuations from separate sTWR integrators, each with their own c parameter and a parameter to determine weight between them. With this 2-integrator sTWR model (employing 3 parameters: c_x, the scalar parameter of one integrator, c_y, the scalar

[1] It is interesting to note that overmatching (behavioral allocation that is more extreme than reinforcement proportions) was observed during phase 1 of all three experiments, despite consistent evidence that concurrent VI-VI schedules typically produce undermatching (Baum, 1979) and overmatching has usually been observed only when switching options incurred significant travel requirements or change-over delays (Baum, 1982). Given that the changeover delay employed was relatively short at 3s, the cause of this overmatching is unclear.

parameter of the other integrator, and W_x, the parameter dictating the relative weight given to the first integrator), we can achieve better performance (R^2 = .981) than the 2-integrator EWMA (which also has 3 parameters), and comparable performance to the 3-integrator EWMA (which has 5 parameters). While this 2-integrator version of sTWR is too theoretically fraught and computationally intensive to seriously consider other than as a hypothetical, it serves to illustrate the potentially misleading success of the multi-integrator EWMA fits. So, while it is possible that these data are truly best *described* by the 3-integrator EWMA model, it is also possible that they may not be best *understood* by this account.

To complicate things, there is another factor that could be affecting the learning rates of our valuation functions: volatility. Here, we define volatility specifically as the variance of reinforcement conditions in the environment. There is limited, but compelling evidence to suggest that volatility modulates learning rates. Behrens and colleagues (2007) found that humans optimally tune learning rates in response to volatility. Saito and colleagues (2014) found that choice data from monkey subjects could be described well by a model which adjusted learning rate in response to volatility. Further, in a recent study with rats, Piet and colleagues (2017), found that subjects optimally adjusted their learning rate in response to volatility. Similarly, Piray and Daw (2021) have proposed a model for learning based on the joint estimation of stochasticity and volatility (based on the idea that optimal decisions in more volatile conditions require higher learning rates) which showed efficacy in simulating both human and animal data. Our analysis employed learning rate parameters that remained constant over the course of the

experiment, so future model development should explore the incorporation of learning rates that vary with volatility.

To sum up, testing data from Experiment 3 challenges the accounts of two primary models under investigation, sTWR and a EWMA. While these data are well-described by more complex models with more parameters, it is still unclear if those models are actually providing an accurate account of recovery observed during testing, or if those models simply have more capability to fit to the test data. Further study will likely clarify this question, but before we can be too confident that these data are best explained by investigation of these functions, we should turn to an equally critical question: whether the formulation of our decision rule is confounding this investigation.

Decision rules

The second theoretical locus of potential responsibility for the form of recovery obtained in these experiments is our decision rule. In all three experiments, at the longer (or longest) delay tested, recovery of choice failed to rise significantly above indifference (.5). Following Experiments 1 and 2, we suspected that a testing effect of some sort may be influencing our data and attenuating recovery—but the between-groups comparison of Experiment 3 failed to show a greater degree of recovery. Together, these three experiments suggest a potential drift toward indifference (or exploration) as test delay increases, possibly because subjects are reverting to a more exploratory or stochastic mode of behavior. Setting aside valuation models for the moment, we will explore how different formulations of our decision rule could produce such a drift.

The decision rule we employed was based on an extension of Herrnstein's matching law (1961). The matching law has been widely studied in diverse choice

scenarios with diverse subjects and experimental methods (for a review of empirical study on the matching law, see Davison & McCarthy, 1988). The formulation we employed follows Baum and Rachlin's (1969) "concatenation" of the matching law—where they posited that value, as a construct, could comprise the product of various parameters of options (like reinforcement rate, magnitude, and immediacy). We then calculated relative preferences over time using the valuation functions discussed above.

While we employed a proportional formulation of the matching law, the drift to indifference may be understood better through using the ratio formulation of Baum's generalized matching law (1974). This version of the matching law set the ratio of behavioral allocation (B_1 / B_2) as equal to the ratio of reinforcement (R_1 / R_2), which is multiplied by a parameter corresponding to bias (b) between options and raised to a parameter (s) representing sensitivity to reinforcement ($B_1 / B_2 = b(R_1 / R_2)^s$). Within this theoretical framework, a drift to indifference could represent a decrease in sensitivity: as sensitivity decreases, the ratio of reinforcers will approach unity, which here equates to equal allocation to the two options, or indifference. By this account, animals could perform indifferent responding (equally allocating to both options) despite preserving relative valuations. In other words, the animals could be tracking valuations by some process like we have described, but the ratio of those valuations may not be the prevailing factor controlling behavioral allocation.

Sensitivity has been theoretically linked to discriminability, following Davison and Tustin's (1999) and Davison & Nevin's (1999) suggestion that matching sensitivity to reinforcement ratios may be better conceptualized as the discriminability of reinforcer contingencies. With this lens, we may view this drift as a decrease in the discriminability

of either the two options themselves (stimulus-response discriminability: d_{sb}, i.e., the extent to which an animal can determine "which response is associated with this stimulus?") or discriminability of the reinforcement schedules associated with the two options (response-reinforcement discriminability: d_{br}, i.e., the extent to which an animal can determine "what is the reinforcement history associated with this response?").

A change in either dimension of discriminability could occur with a variety of changes that occur during the testing phase. During the test delay, the options are not available to the subjects in their home cages, and when the options are finally made available, they are not baited. The dramatic drop in overall reinforcer rate during this time could have some influence on discriminability, as is suggested by a study by Bizo and White (1994) where rats shifted behavioral allocation to reflect the current distribution of reinforcement more slowly with lower overall rates of reinforcement. A decrease in discriminability could also simply result from the passage of time (White, 2002), potentially due to psychophysical properties of the remembering process.

In contrast to this decision rule from matching theory, the paradigm of many explore/exploit models hypothesizes that a softmax function guides the decision to exploit (choose the option with the highest value) or explore (choose another option) (Daw et al., 2006)—and the influence of the highest valued option is determined by an inverse temperature or "softmax gain" parameter (Addicott et al., 2017). In the context of Experiments 1, 2, and 3, the time delay could be somehow linked to a decrease in this temperature parameter, meaning the observed trend toward indifference is a consequence

of subjects returning to a mode of exploring both options, despite preserving differences[2] in valuation between the options. As a caveat, this interpretation would be somewhat inconsistent with previous findings where after a long delay, despite showing indifference between two previously reinforced patches, subjects almost never chose a third patch that was never reinforced or extinguished—suggesting that "exploration" would at least be limited only to previously reinforced patches. (Devenport & Devenport, 1993; Devenport et al., 2005). Further analysis will be needed to evaluate the efficacy of this alternative.

Finally, while there is limited evidence showing the effect of volatility on variable learning rates in our valuation functions, it is also possible that volatility might influence the sensitivity/discriminability or inverse temperature of our decision rule. This notion follows the idea that when animals are in highly variable environments, there is an adaptive value in tending toward exploration (or indifference), which would entail more sampling of the various options to determine their uncertain value. In agreement with this idea, there is some evidence linking volatility to greater exploration, as quantified by a lower inverse temperature parameter (Knox et al., 2011; Wang et al., 2023). It is unclear whether volatility itself would produce an effect on discriminability, or whether volatility may simply moderate the effects of test delay or reinforcement rate on discriminability.

In support of this potential effect, there is some evidence that exposure to variable reinforcement conditions decreases sensitivity over time in a steady-state experiment (Todorov et al., 1983) and over the course of a series of experimental blocks with different reinforcement conditions (McLean et al., 2018)—but volatility's effect on

[2] Note that decision rules based on the difference between option valuations, such as the softmax functions employed in many explore/exploit models, have poorer empirical support than models employing decision rules based on the ratio of these valuations (Worthy, 2008; Gibbon, 1977).

sensitivity for various lengths of test delay is essentially unexplored. Additionally, in several of the studies that examined the effect of volatility on valuation (Behrens et al., 2007; Saito et al., 2014, Piet et al., 2017), it is possible that the appearance of some changes in learning rates could also be produced by variation of the sensitivity of the decision rule employed, but further study is needed to explore this question directly.

Future directions

The potential influence of reinforcement rate and recency on sensitivity or discriminability may be readily evaluated with a simple control experiment: replicating Experiment 3, but omitting phase 2, in which reinforcement conditions are reversed. If preference for A remains stable for all of the four test delays, we can assert that the discriminability of the two options is not specifically linked to either the drop in reinforcement rate or the passage of time alone. If preference does show a drift toward indifference, further manipulation could distinguish the effects of time passing and the decrease in reinforcement rate.

Moreover, if such an experiment does allow us to rule out reinforcement rate and recency as the determining influence on sensitivity, Experiment 3 could also be adapted to examine the influence of volatility. If phase 2 is moved instead to the middle of phase 1 (i.e., seven days favoring option A, followed by 2 days favoring option B, followed by an additional seven days favoring option A), a trend toward indifference as test delay increases could be tentatively linked to the volatile history of reinforcement conditions. Further analysis would be required to determine whether volatility's effect is best conceptualized as an influence on learning rate or an influence on the decision rule, but

experimentally isolating and theoretically incorporating the effect of volatility could help to greatly improve our understanding moving forward.

Finally, in attempting to determine why spontaneous recovery of choice differed from previous studies of the topic, timescale stands out as a potentially crucial difference. However, very few studies have examined dynamic averaging or spontaneous recovery of choice on different time scales. The longest delay condition used by Devenport and colleagues (1997) in manipulating test delay to produce spontaneous recovery was 24 hours. A study with honeybees found some limited evidence of preference reversals consistent with TWR within a single day (Cheng, 2012), but once the bees had slept, they simply preferred the most recently reinforced option. A future study could evaluate the performance of these models on various timescales to determine if circadian patterns similarly influence dynamic averaging processes for rats.

REFERENCES

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. *Neuropsychopharmacology*, *42*(10), 1931-1939.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *J Exp Anal Behav*, *22*(1), 231-242.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *J Exp Anal Behav*, 32(2), 269-281.

Baum, W. M. (1982). Choice, changeover, and travel. *J Exp Anal Behav, 38*(1), 35-49.

Baum, W. M., & Rachlin, H. C. (1969). Choice as time allocation. *J Exp Anal Behav*, 12(6), 861-874.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214-1221.

Bizo, L. A., & White, K. G. (1994). The behavioral theory of timing: Reinforcer rate determines pacemaker rate. *J Exp Anal Behav*, 61(1), 19-33.

Brooks, D. C., & Bouton, M. E. (1993). A retrieval cue for extinction attenuates spontaneous recovery. *J Exp Psychol Anim Behav Process,* 19(1), 77-89.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychol Rev*, *58*(5), 313-323.

Cramer, J. S. (2003). The origins and development of the logit model. In *Logit Models from Economics and Other Fields* (pp. 149-157). Cambridge University Press.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature, 268*(5616), 137-139.

Cheng, K. (2012). Testing Mathematical Laws of Behavior in the Honey Bee. In *Honeybee Neurobiology and Behavior* (pp. 457-470). Dordrecht: Springer Netherlands.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441: 876–879.

Davison, M. & McCarthy, D. (1988). *The matching law: A research review*. Hillsdale, NJ: Erlbaum.

Davison, M., & Jenkins, P. E. (1985). Stimulus discriminability, contingency discriminability, and schedule performance. *Animal Learning & Behavior*, *13*(1), 77-84.

Davison, M., & Nevin, J. (1999). Stimuli, reinforcers, and behavior: an integration. *J Exp Anal Behav*, 71(3), 439-482.

Devenport, L. D. (1998). Spontaneous recovery without interference: Why remembering is adaptive. *Animal Learning & Behavior, 26*(2), 172-181.

Devenport, J. A., & Devenport, L. D. (1993). Time-dependent decisions in dogs (Canis familiaris). *Journal of Comparative Psychology, 107*(2), 169-173.

Devenport, L.D., & Devenport, J. A. (1994). Time-dependent averaging of foraging information in least chipmunks and golden-mantled ground squirrels. *Animal Beh, 47*(4), 787-802.

Devenport, L., Hill, T., Wilson, M., & Ogden, E. (1997). Tracking and averaging in variable environments: A transition rule. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*(4), 450-460.

Devenport, J. A., Patterson, M. R., & Devenport, L. D. (2005). Dynamic averaging and foraging decisions in horses (Equus callabus). *J Comp Psychol*, *119*(3), 352-358.

Dow, S. M., & Lea, S. E. G. (1987). Foraging in a changing environment: Simulations in the operant laboratory. In M. L. Commons, A. Kacelnik, & S. J. Shettleworth (Eds.), *Quantitative analyses of behavior, Vol. 6. Foraging.* Lawrence Erlbaum Associates, Inc.

Ellson, D. G. (1938). Quantitative studies of the interaction of simple habits. I. Recovery from specific and generalized effects of extinction. *J of Exp Psychology*, *23*(4), 339-358.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychol Rev*, *62*(3), 145-154.

Fleshler, M., & Hoffman, H. S. (1962). A progression for generating variable-interval schedules. *Journal of the Experimental Analysis of Behavior, 5,* 529-530.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279-325.

Grant, D. A., Hunter, H. G., & Patel, A. S. (1958). Spontaneous recovery of the conditioned eyelid response. *J Gen Psychol*, *59*(1), 135-141.

Haberlandt, K., Hamsher, K., & Kennedy, A. W. (1978). Spontaneous recovery in rabbit eyelid conditioning. *J Gen Psychol*, *98*(2d Half), 241-244.

Harley, C. B. (1981). Learning the evolutionarily stable strategy. *J Theor Biol, 89*(4), 611-633.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *J Exp Anal Behav*, *4*, 267-272.

Iigaya, K., Ahmadian, Y., Sugrue, L. P., Corrado, G. S., Loewenstein, Y., Newsome, W. T. et al. (2019). Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nat Commun*, *10*(1), 1466.

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *J of Mathematical Psychology*, *66*, 59-69.

Killeen, P. R. (1981). Averaging Theory. In C. M. Bradshaw, E. S. Szabadi, & C. F. Lowe (Eds), *Quantification of Steady-State Operant Behavior* (pp. 21-34). New York: Elsevier.

Klapes, B., Riley, S., & McDowell, J. J. (2018). Toward a contemporary quantitative model of punishment. *J Exp Anal Behav*, *109*(2), 336-348.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2011). The nature of belief-directed exploratory choice in human decision-making. *Front Psychol,* 2, 398.

Lattal, K. M., Mullen, M. T., & Abel, T. (2003). Extinction, renewal, and spontaneous recovery of a spatial preference in the water maze. *Behav Neurosci*, *117*(5), 1017-1028.

Leising, K. J., Wong, J., & Blaisdell, A. P. (2015). Extinction and spontaneous recovery of spatial behavior in pigeons. *J Exp Psychol Anim Learn Cogn*, *41*(4), 371-377.

Lea, S. E., & Dow, S. M. (1984). The integration of reinforcements over time. *Ann N Y Acad Sci, 423,* 269-277.

Mazur, J. E. (1995). Development of preference and spontaneous recovery in choice behavior with concurrent variable-interval schedules. *Animal Learning & Behavior, 23*(1), 93–103.

Mazur, J. E. (1996). Past experience, recency, and spontaneous recovery in choice behavior. *Animal Learning & Behavior, 24*(1), 1-10.

McLean, A. P., Grace, R. C., Shevchouk, O. T., & Cording, J. R. (2018). Rat choice in rapidly changing concurrent schedules. *J Exp Anal Behav*, *109*(2), 313-335.

McNamara, J., & Houston, A. (1980). The application of statistical decision theory to animal behaviour. *J Theor Biol, 85*(4)*,* 673-690.

McNamara, J. M., & Houston, A. I. (1987). Memory and the efficient use of information. *J Theor Biol, 125*(4), 385-395.

Myerson, J., & Miezin, F. M. (1980). The kinetics of choice: An operant systems analysis. *Psychological Review, 87*(2), 160-174.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cogn Psychol*, *85*, 43-77.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

Kacelnik, A. (1979). Studies of foraging behaviour and time budgeting in great tits (parus major) [PhD thesis]. University of Oxford.

Pavlov, I.P. 1927. *Conditioned reflexes*. Oxford University Press, Oxford.

Piet, A., Hady, A. E., & Brody, C. D. (2017). Rats optimally accumulate and discount evidence in a dynamic environment. *arXiv*, 1710.05945v1.

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nat Commun,* 12(1), 6587.

Pyke, G.H. (1984). Optimal foraging theory: A critical review. *Ann Review of Ecology and Systematics 15*, 523–575.

Quirk, G. J. (2002). Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery. *Learn Mem*, *9*(6), 402-407.

Rescorla, R. A. (2004). Spontaneous recovery. *Learn Mem*, *11*(5), 501-509.

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Robbins, S. J. (1990). Mechanisms underlying spontaneous recovery in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*(3), 235-249.

Saito, H., Katahira, K., Okanoya, K., & Okada, M. (2014). Bayesian deterministic decision making: a normative account of the operant matching law and heavy-tailed reward history dependency of choices. *Front Comput Neurosci*, *8*, 18.

Shahan, T. A., & Craig, A. R. (2017). Resurgence as Choice. *Beh Processes, 141*, 100-127.

Skinner, B.F. (1938) *The Behavior of Organisms: An Experimental Analysis.* B.F. Skinner Foundation, Cambridge.

Stephens, D. W., & Dunlap, A. S. (2017). Foraging. In *Learning and Memory: A Comprehensive Reference* (pp. 237-253). Elsevier.

Su, C.-T., & Hsu, C.-C. (2004). On-line tuning of a single EWMA controller based on the neural technique. *International Journal of Production Research*, *42*(11), 2163-2178.

Todorov, J. C., de Oliveira Castro, J. M., Hanna, E. S., Bittencourt de Sa, M. C., & Barreto, M. Q. (1983). Choice, experience, and the generalized matching law. *J Exp Anal Behav*, *40*(2), 99-111.

Wang, S., Gerken, B., Wieland, J. R., Wilson, R. C., & Fellous, J. M. (2023). The effects of time horizon and guided choices on explore-exploit decisions in rodents. *Behav Neurosci*.

White, K. G. (2002). Psychophysics of Remembering: The Discrimination Hypothesis. *Current Directions in Psychological Science*, *11, No. 4*(Aug.), 141-145.

Worthy, D. A., Maddox, W. T., & Markman, A. B. (2008). Ratio and difference comparisons of expected reward in decision-making tasks. *Mem Cognit*, *36*(8), 1460-1469.

Zhang, R., Gong, W., Grzeda, V., Yaworski, A., & Greenspan, M. (2013). An Adaptive Learning Rate Method for Improving Adaptability of Background Models. *IEEE Signal Processing Letters*, *20*(12), 1266-1269.

FIGURES

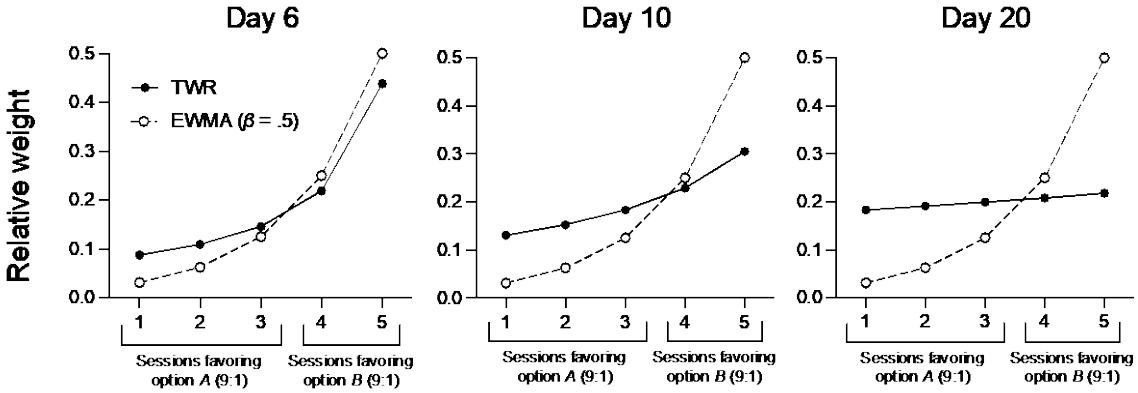

Figure 1. Relative weights assigned to past experiences over time by TWR and a EWMA model. Weights of 5 consecutive daily sessions followed by no new experiences are displayed at 3 future time points: day 6 (the day after the final session), day 10 (5 days after the final session), and day 20 (15 days after the final session).

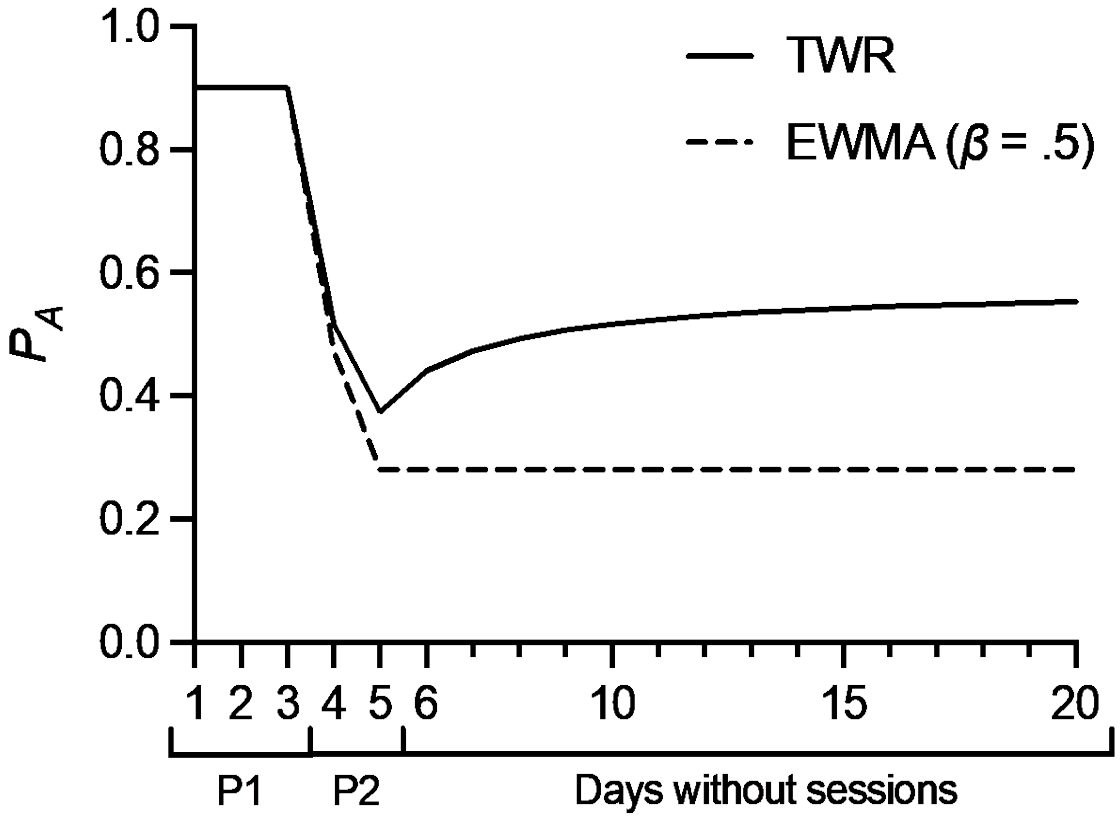

Figure 2. Predictions of preference for option A (P_A) as a function of time (days) from TWR and a EWMA model in a spontaneous recovery of choice procedure. P_A is calculated as the relative value of option A ($P_A = V_A / (V_A + V_B)$). P1 indicates sessions when option A was reinforced at a ratio of 9:1 (relative to option B) while P2 indicates sessions when option B was reinforced at a ratio of 9:1 (relative to option A).

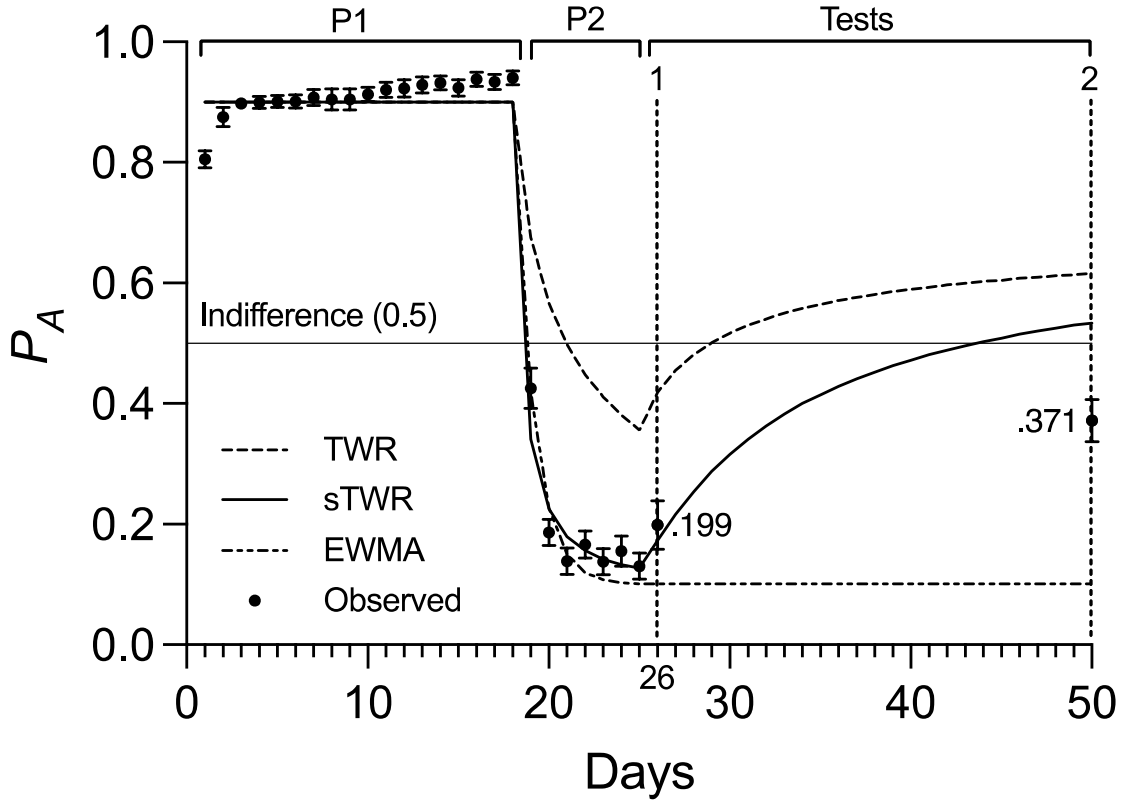

Figure 3. Preference for option A (P_A) as a function of time (days) during Experiment 1. Phases 1 and 2 are indicated by P1 and P2, respectively, and tests are labeled above the data. Error bars represent the standard error of the mean. Predictions from TWR, sTWR (c = 2.27), and a EWMA model (β = .603) are based on programmed reinforcement rates.

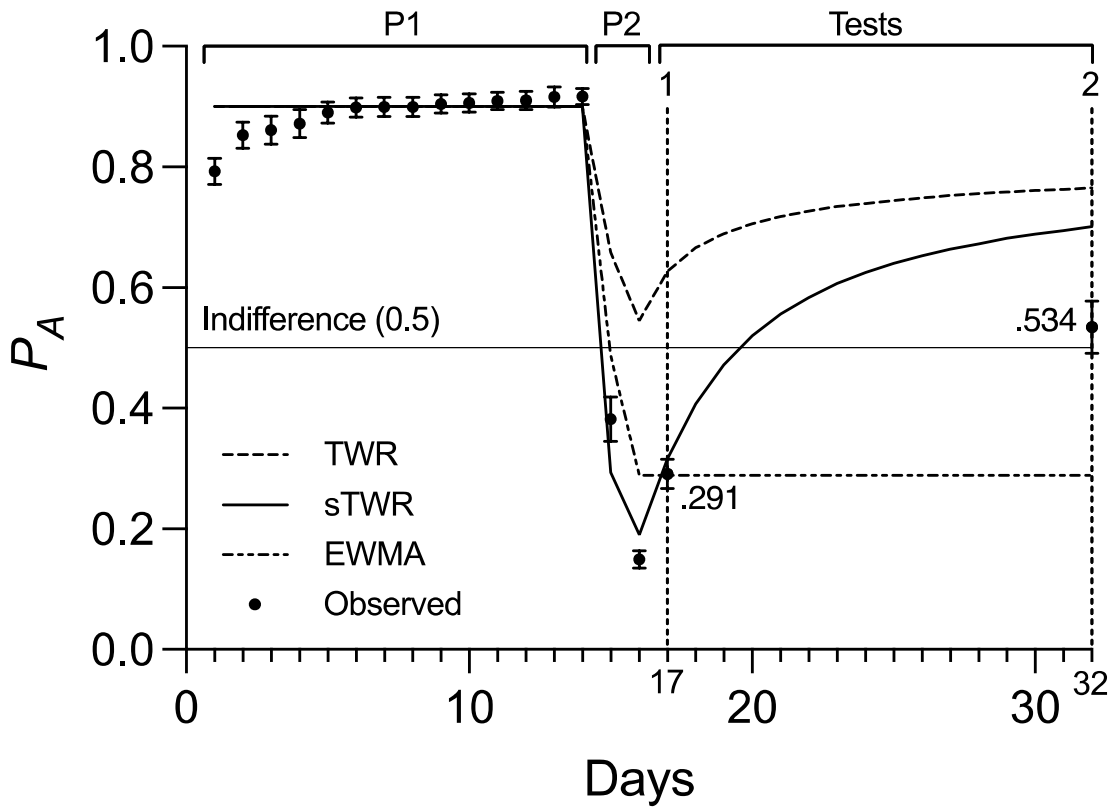

Figure 4. Preference for option A (P_A) as a function of time (days) during Experiment 2. Phases 1 and 2 are indicated by P1 and P2, respectively, and tests are labeled above the data. Error bars represent the standard error of the mean. Predictions from TWR, sTWR (c = 2.53), and a EWMA model (β = .514) are based on programmed reinforcement rates.

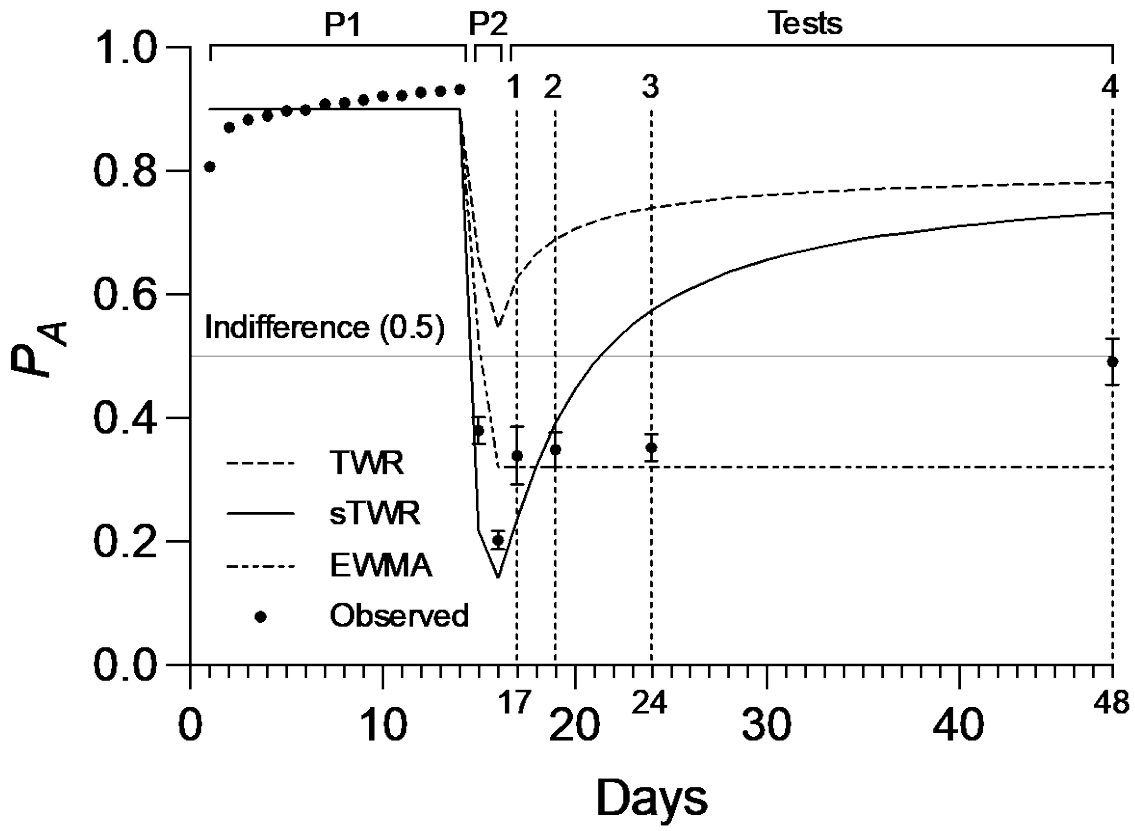

Figure 5. Preference for option A (P_A) as a function of time (days) in Experiment 3. Phases 1 and 2 are indicated by P1 and P2, respectively, and tests are labeled above the data. Predictions from sTWR use the best fit parameter value of $c = 3.15$. Predictions from a EWMA model use the best fit parameter value of $\beta = .47$. Test sessions occurred on days 17, 19, 24, and 48, corresponding to delays of 1, 3, 8, and 32 days, respectively.

TABLES

Table 1
Experiment 3 Test Session Data

Test delay	Day	Proportional Preference for A	*SE*	Logit Preference for A	*SE*
1	17	.339	.047	-.763	.261
3	19	.357	.034	-.620	.156
8	24	.359	.037	-.601	.159
32	48	.491	.038	-.040	.155

Table 2
Regression of Test Delay's effect on Logit Preference in Experiment 3

	Estimate	*SE*	*95% CI*		*p*
			LL	*UL*	
Intercept	-0.747	.121	-0.992	-0.502	<.001***
Test Delay	0.022	.007	.007	.037	.004**

Table 3
Experiment 3 Model Fit Comparison

Model	Parameter fit	*RSS*	R^2	*AIC*	*BIC*
TWR	*none*	0.643	.513	-66.74	-65.75
sTWR	$c = 3.15$	0.164	.876	-94.09	-93.09
sTWR	$c_x = 28.30$ $c_y = 1.23$ $W_x = .552$	0.023	.981	-129.61	-126.62
EWMA	$\beta = 0.476$	0.079	.940	-108.71	-107.71
EWMA (2 integrator)	$\beta_x = .823$ $\beta_y = .382$ $W_x = .630$	0.028	.979	-125.76	-122.77
EWMA (3 integrator)	$\beta_x = .921$ $\beta_y = .00053$ $\beta_z = .437$ $W_x = .483$ $W_y = .00000028$	0.015	.988	-133.54	-128.56