Received: 26 July 2023 | Accepted: 9 February 2024

DOI: 10.1002/jeab.908

RESEARCH ARTICLE

Journal of the Experimental Analysis of Behavior SEAB

Influence of reinforcement and its omission on trial-by-trial changes of response bias in perceptual decision making

Maik C. Stüttgen[1] | **Andrea Dietl**[1] | **Vanya V. Stoilova Eckert**[1] | **Luis de la Cuesta-Ferrer**[1] | **Jan-Hendrik Blanke**[1] | **Christina Koß**[2] | **Frank Jäkel**[2]

[1]Institute of Pathophysiology, University Medical Center of the Johannes Gutenberg University Mainz, Germany

[2]Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Germany

Correspondence
Maik C. Stüttgen, Institute of Pathophysiology, University Medical Center of the Johannes Gutenberg University Mainz, 55128 Mainz, Germany.
Email: maik.stuettgen@uni-mainz.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: JA 1878/2-1, STU 544/6-1

Editor-in-Chief: Suzanne H. Mitchell
Handling Editor: Sarah Cowie

Abstract

Discrimination performance in perceptual choice tasks is known to reflect both sensory discriminability and nonsensory response bias. In the framework of signal detection theory, these aspects of discrimination performance are quantified through separate measures, sensitivity (d') for sensory discriminability and decision criterion (c) for response bias. However, it is unknown how response bias (i.e., criterion) changes at the single-trial level as a consequence of reinforcement history. We subjected rats to a two-stimulus two-response conditional discrimination task with auditory stimuli and induced response bias through unequal reinforcement probabilities for the two responses. We compared three signal-detection-theory-based criterion learning models with respect to their ability to fit experimentally observed fluctuations of response bias on a trial-by-trial level. These models shift the criterion by a fixed step (1) after each reinforced response or (2) after each nonreinforced response or (3) after both. We find that all three models fail to capture essential aspects of the data. Prompted by the observation that steady-state criterion values conformed well to a behavioral model of signal detection based on the generalized matching law, we constructed a trial-based version of this model and find that it provides a superior account of response bias fluctuations under changing reinforcement contingencies.

KEYWORDS
criterion, rat, response bias, reward, signal detection theory

Signal detection theory (SDT) constitutes a widely adopted framework for modeling perceptual decisions in psychophysical tasks (for review, see Green & Swets, 1988, and Hautus et al., 2022). Signal detection theory breaks down the experimentally observed discrimination performance into two independent performance indices representing sensitivity and response bias. The sensitivity measure d' quantifies the degree to which the two stimuli lead to psychophysically discriminable sensations for a given subject. The bias measure (β or criterion c) quantifies the degree to which a subject emits one response more frequently than the other. In psychophysics, response bias is usually treated as a nuisance factor, and d' is therefore used as bias-free index of perceptual ability. However, the study of bias is interesting in its own right—for example, to test some of SDT's core assumptions such as the shape of the receiver operating characteristic (ROC) curve (Swets, 1961a, 1961b), to investigate mechanisms underlying perceptual learning (Gold & Ding, 2013), and to examine nonsensory factors that influence sensory-guided choices (Alsop, 1998).

There are two well-established procedures for experimentally manipulating response bias in perceptual choice tasks—namely, using unequal stimulus presentation probabilities (e.g., presenting Stimulus 1 more often than Stimulus 2) and using unequal payoffs (e.g., providing reinforcement more often for correct choices of one stimulus category than for the other). Neither stimulus presentation probabilities nor the payoff matrix is usually

made explicit for the subject, but both can be estimated based on the recent history of stimuli, choices, and outcomes. Importantly, SDT itself does not specify how subjects adapt their criterion to a certain experimental situation (criterion learning). Although several models have been proposed as to how subjects may shift their criterion after feedback (e.g., Boneau & Cole, 1967; Busemeyer & Myung, 1992; Dorfman & Biderman, 1971; Erev, 1998; Funamizu, 2021; Kac, 1962; Lak et al., 2017; Lak, Okun, et al., 2020; Luce, 1963; Mill et al., 2014; Stüttgen, Yildiz, et al., 2011; Treisman & Williams, 1984), none of these models has been subjected to extensive experimental scrutiny.

Previous research has shown that a simple income-based criterion learning model is able to fit results from different experiments with rats, pigeons, and mice (Stoilova et al., 2020; Stüttgen et al., 2013; Vandevelde et al., 2023). Here, we test the ability of this model to fit experimental results obtained with rat subjects in two different experiments and compare its performance with that of two related models of criterion learning. In the remainder of the Introduction, we will first explain the concept of adaptive criterion setting within the SDT framework and then introduce the three criterion learning models and describe the design of the two experiments.

Criterion setting in the SDT framework

We first briefly review how decisions are made within the SDT framework (for a detailed outline, see Hautus et al., 2022). We consider the situation where an observer is performing a two-stimulus two-response conditional discrimination task. This procedure is also referred to as "yes/no," "single-interval forced choice," or simply "single-interval" task and should not be confused with the two-alternative forced-choice task (Stüttgen, Schwarz, et al., 2011; Wichmann & Jäkel, 2018). Signal detection theory posits that each presentation of a stimulus gives rise to a random variable X on a decision axis, where X is drawn from one of two equal-variance normal distributions that correspond to the two stimuli, S1 and S2 (Figure 1A). The observer decides on the probable identity of the currently perceived stimulus based on a comparison of X with a decision criterion c. If $X < c$, the observer will respond "S1" (emit R1), and if $X > c$, the observer will respond "S2" (emit R2). The distance between the means of the two distributions determines the degree to which a given value of X is informative as to the identity of the distribution from which it has been drawn. The distance between the two means divided by their standard deviation is denoted d' and constitutes an index pertaining to the discriminability of the two stimuli. For the example shown in Figure 1A, $d' = 2$.

The actual discrimination performance (i.e., the proportions of correct and incorrect choices in S1 and S2 trials) results from the combination of d' and the location of

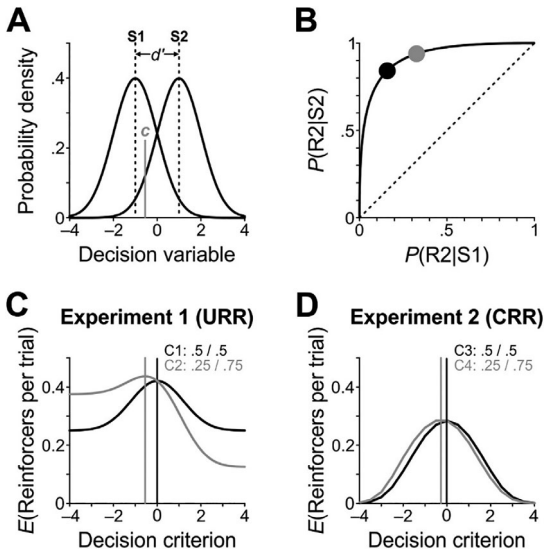

FIGURE 1 Illustration of criterion setting in signal detection theory. (A) The two Gaussian distributions correspond to the probability densities of two different stimuli (S1 and S2) on a decision axis (a.k.a. evidence variable). The criterion c is located at −0.55 (gray vertical line); thus, the observer is biased toward R2. (B) ROC curve for $d' = 2$. The black dot marks a decision criterion of 0, and the gray dot marks $c = -0.55$, as shown in panel A. (C) Sample objective reward functions (representing the total expected probability of reinforcement in a trial dependent on the criterion) for $d' = 2$ and two different sets of reinforcement probabilities (Conditions C1 and C2, black and gray curves, respectively) in Experiment 1 (uncontrolled reinforcement ratio [URR]). Optimal criterion values are represented by vertical lines. (D) As in panel C, but for Experiment 2 (controlled reinforcement ratio [CRR]) and conditions C3 and C4.

the criterion. If c is located exactly halfway between the two means (i.e., if $c = 0$), then the percentages of correct S1 and correct S2 trials will be identical (for $d' = 2$, both are 84%). In Figure 1A, however, $c = -0.55$, so the observer is biased toward R2. Such a bias can be induced by employing unequal stimulus presentations probabilities (here, presenting S2 more often than S1) or by providing reinforcement more frequently in correct S2 than in correct S1 trials. Inducing response biases of different magnitudes will yield pairs of hit rates—here, R2 on S2 trials, $P(R2|S2)$—and false-alarm rates—here, R2 on S1 trials, $P(R2|S1)$. Signal detection theory predicts that these pairs of hit and false alarm rates will all be located on an ROC or "isosensitivity" curve—that is, a curve containing all possible pairs of hit and false-alarm rates for a given d' (Figure 1B).

Three SDT-based models of criterion learning

There is ample evidence suggesting that the criterion is not stationary but affected by stimuli, choices, and outcomes of the immediately preceding trials (e.g., Benjamin et al., 2009; Stoilova et al., 2020; Stüttgen, Yildiz, et al., 2011; Stüttgen et al., 2013). The mechanisms

underlying this gradual adaptation of the decision criterion are unknown.

Assuming that the subject has sufficient experience with the two stimuli to estimate their corresponding distributions, it is reasonable to propose that the criterion is shifted following feedback. In a scenario where correct responses are reinforced and incorrect responses are of no consequence, the simplest criterion learning model would entail shifting the criterion to the left after a correct S2 → R2 trial (thus increasing the chance of R2 in the next trial) and vice versa after a correct S1 → R1 trial. This mechanism was first proposed by Dorfman and Biderman (1971) but eventually discarded because the criterion in this model quickly runs off from zero and eventually produces exclusive choice of one response (which happens because of this model's inherent positive feedback loop). However, this problem can be solved by postulating that the criterion on a trial is an exponentially weighted average of all previous criteria (Stüttgen et al., 2013). In this model (henceforth, "Model 1" or "income-based model"), the criterion updates according to the following equation:

$$c(t+1) = \gamma \times c(t) + \delta \times (\mathrm{Rf}_{\mathrm{R1}} - \mathrm{Rf}_{\mathrm{R2}}).$$

Here, $c(t)$ is the criterion on trial t; γ is a leak factor restricted to range from 0 to 1, which pulls the criterion back toward 0 (the midpoint between the two stimulus distributions) and thus constitutes a leaky integration mechanism; δ is a learning-rate parameter, which determines the step size of the criterion adjustment; and $\mathrm{Rf}_{\mathrm{R1}}$ and $\mathrm{Rf}_{\mathrm{R2}}$ correspond to reinforcement for R1 or R2, respectively, and can take values of either 0 or 1. Thus, the criterion value is incremented by δ if R1 was reinforced and decremented by δ if R2 was reinforced. In trials without reinforcement, both $\mathrm{Rf}_{\mathrm{R1}}$ and $\mathrm{Rf}_{\mathrm{R2}}$ are 0, so the criterion is pulled back toward 0 to an extent determined by γ.

However, it is equally conceivable that learning is actually driven by failure to obtain reinforcement (reinforcement omission). In fact, adjusting the criterion (exclusively) on error trials is a mechanism widely believed to hold true in human psychophysics because it predicts the frequently observed phenomenon of probability matching (Dorfman, 1969; Dorfman & Biderman, 1971; Dorfman et al., 1975; Friedman et al., 1968; Killeen et al., 2018; Thomas, 1975). So, we conceived of another model (Model 2), which learns exclusively on nonreinforced trials (which not only include all error trials but, in our experiment, also correct trials in which reinforcement is omitted):

$$c(t+1) = \gamma \times c(t) + \upsilon \times (\mathrm{noRf}_{\mathrm{R2}} - \mathrm{noRf}_{\mathrm{R1}}).$$

The learning rate parameter for this model is denoted υ (upsilon), and $\mathrm{noRf}_{\mathrm{R2}}$ and $\mathrm{noRf}_{\mathrm{R1}}$ represent nonreinforced R2 and R1 trials, respectively, and take a value of 1 when no reinforcement occurs and 0 otherwise. On trials with reinforcement, the term on the right-hand side is 0; thus, the criterion is pulled toward 0 at a rate determined by γ.

Finally, Model 3 combines learning on both reinforced and non-reinforced trials:

$$c(t+1) = \gamma \times c(t) + \delta \times (\mathrm{Rf}_{\mathrm{R1}} - \mathrm{Rf}_{\mathrm{R2}}) + \upsilon \times (\mathrm{noRf}_{\mathrm{R2}} - \mathrm{noRf}_{\mathrm{R1}}).$$

Description of the experiments

To arbitrate between the three models, we conducted two experiments with rat subjects performing a two-stimulus two-choice conditional discrimination task. Response bias was manipulated by implementing unequal reinforcement probabilities for the two responses, as illustrated in Figure 1. Reinforcement contingencies were unsignaled and changed every five sessions (see Methods for details). The main difference between the experiments was the employed schedules of reinforcement. Experiment 1 featured an uncontrolled reinforcer ratio schedule in which reinforcers were delivered with a certain probability after correct responses—for example, $P(\mathrm{Rf}|\mathrm{S1,R1}) = P(\mathrm{Rf}|\mathrm{S2,R2}) = .50$—and reinforcers were never delivered after incorrect responses. In this situation, the relative probabilities of obtaining reinforcement for R1 or R2 depend not only on the programmed probabilities but also on the behavior of the subject. For example, a subject with $d' = 2$ and $c = 0$ will, in the long run, obtain the same number of reinforcers for both R1 and R2, so both the reinforcement ratio $\mathrm{Rf}_{\mathrm{R1}}/\mathrm{Rf}_{\mathrm{R2}}$ and the response ratio R1/R2 will be 1. However, a subject with a criterion value of −0.55 as in Figure 1A–B will emit R2 more often than R1 (in 63% of all trials) and therefore also obtain more reinforcers after R2 than after R1 (in 67% of S1 trials and 94% of S2 trials, so 58% of all reinforcers are produced by R2). So, in this example, the programmed $\mathrm{Rf}_{\mathrm{R1}}/\mathrm{Rf}_{\mathrm{R2}}$ is 1, but the obtained $\mathrm{Rf}_{\mathrm{R1}}/\mathrm{Rf}_{\mathrm{R2}}$ is 0.72, and the response ratio R1/R2 is 0.58.[1]

In Experiment 2, a controlled reinforcer ratio schedule was employed. This schedule differs from the uncontrolled reinforcer ratio schedule in the way that reinforcement is allocated to correct responses (see Methods for details). Briefly, the next reinforcement is assigned to either of the two responses with a certain probability (here, .25 for R1 and .75 for R2, a

[1]The probability of R1 is $P(\mathrm{R1}) = P(\mathrm{R1}|\mathrm{S1}) \times P(\mathrm{S1}) + P(\mathrm{R1}|\mathrm{S2}) \times P(\mathrm{S2})$. With $d' = 2$, $c = -0.55$, $P(\mathrm{S1}) = P(\mathrm{S2}) = 0.5$, and $P(\mathrm{Rf}|\mathrm{R1,S1}) = P(\mathrm{Rf}|\mathrm{R2,S2}) = 0.5$, it follows that $P(\mathrm{R1}) = (0.67 \times 0.5) + (0.06 \times 0.5) = 0.37$, and likewise $P(\mathrm{R2}) = P(\mathrm{R2}|\mathrm{S1}) \times P(\mathrm{S1}) + P(\mathrm{R2}|\mathrm{S2}) \times P(\mathrm{S2}) = (0.33 \times 0.5) + (0.94 \times 0.5) = 0.63$. Thus, the response ratio R1/R2 is $P(\mathrm{R1}) / P(\mathrm{R2}) = 0.37 / 0.63 = 0.58$. The overall probability of reinforcement for the two responses is given by $\mathrm{Rf}_{\mathrm{R1}} + \mathrm{Rf}_{\mathrm{R2}} = P(\mathrm{Rf}|\mathrm{R1,S1}) \times P(\mathrm{R1,S1}) + P(\mathrm{Rf}|\mathrm{R2,S2}) \times P(\mathrm{R2,S2}) = (0.5 \times 0.67 \times 0.5) + (0.5 \times 0.94 \times 0.5) = 0.40$, and the reinforcer ratio $\mathrm{Rf}_{\mathrm{R1}} / \mathrm{Rf}_{\mathrm{R2}}$ is therefore 0.17 / 0.24 = 0.72 (slight inaccuracies due to rounding to the second decimal place).

reinforcement ratio of 1:3), and then a variable number of correct responses (VR) of that type must be completed before that reinforcement is provided, after which the procedure starts over again. The upshot of this procedure is that the reinforcement ratio is fixed—that is, it does not depend on the response allocation of the animal (McCarthy & Davison, 1984; Stubbs & Pliskoff, 1969). As a result, optimal criterion locations and objective reward functions (see Methods for details) differ between experiments even for the same pair of programmed reinforcement probabilities (Figure 1C–D).

METHODS

Subjects

Subjects were four male Long-Evans rats (Janvier Labs), aged 8 weeks and weighing 200–250 g at the start of behavioral training. The animals were housed in a common cage in a ventilated temperature- and humidity-controlled cabinet with an inverted day-night cycle (lights off from 8 a.m. until 8 p.m.). Food was available ad libitum in the home cage throughout the entire experiment. Water was freely available on weekends only. During weekdays, access to water was restricted to behavioral testing. The rats' weight was measured before and after each testing session. Despite water restriction, the animals consistently gained weight over the course of the experiments. All procedures were approved by local authorities (Landesuntersuchungsamt Rheinland-Pfalz) and conducted in agreement with German law as well as directive 2010/63/EU of the European Parliament.

Apparatus and stimuli

Behavioral testing was conducted in a standard operant chamber (ENV-008, Med Associates) measuring 48 × 27 × 28 cm (L × W × H). The chamber was housed in a sound-attenuating wooden cubicle (length, width, and height, all 80 cm). One side wall featured three nose ports (LIC.80117RM, Lafayette Instruments), which allowed us to detect nose entry and to deliver small amounts of water. Each reinforcement amounted to 30 µL of water that was delivered by 0.5 s of pump activation. Houselights provided constant dim illumination but were turned off briefly during timeouts (see below). The experimental hardware was controlled from a PC running custom-written software written in Spike2 via a power 1401 AD converter (Cambridge Electronic Design).

The auditory stimuli were composed of bandpass-filtered white noise bursts with either of two different center frequencies (4096 or 16384 Hz for Stimulus 1 and 2, respectively) and durations of 70 ms. Initial training was conducted with easily discriminable stimulus pairs (±0.4 octaves). As the rats grew more proficient, the bandwidths were gradually increased until the animals performed correctly in about 80% of trials with no further improvement (final bandwidths ranged from ±2.8 to ±3.0 octaves, adjusted individually for each subject). White noise was generated at a sampling rate of 200 kHz and filtered in Matlab (The Mathworks). The resulting vectors were imported into Spike2 (Cambridge Electronic Design, Inc.) and output at the same sampling frequency from the analog output port of the AD converter. The sounds were amplified and presented through a loudspeaker that was attached to the ceiling of the sound-attenuating cubicle. The sound pressure levels of all stimuli were adjusted to 70 dB SPL and calibrated with a ¼ microphone (Microtech Gefell).

Procedure

The rats were trained on a two-stimulus two-response conditional discrimination procedure (for an outline of a single trial, see Figure A1, panel A). The rats could initiate trials by poking into the center port continuously for 400 ms. On each trial, one of two stimuli (S1 and S2) was presented and animals were required to maintain nose poking until stimulus offset. The rats indicated their choice by poking either of the two side ports. A poke into the right choice port (R1) was considered correct following S1, and a poke into the left choice port (R2) was considered correct following S2. Correct responses were reinforced according to the probabilistic schedules (see below). Correct but nonreinforced responses terminated the trial. A new trial could be initiated immediately. Incorrect responses were punished with a timeout of 4 s during which the houselight was turned off. The stimulus sequence was pseudorandomized by concatenating independent sets of 20 trials comprising 10 S1 and 10 S2 trials after shuffling. Within each set, a certain fraction of S1 and S2 trials was assigned reinforcement (if followed by the response designated as correct), corresponding to the reinforcement probability for a given stimulus in a given condition.

If the rats broke fixation during the 400 ms of trial initiation or during stimulus presentation, the trial was counted as a premature response and aborted. Premature responses were punished with a timeout of 4 s, and aborted trials were not repeated. Each session lasted 45 min and contained a median of 551 trials (Experiment 1) and 720 trials (Experiment 2). After 12 weeks of daily training on the task, all rats reliably performed hundreds of trials per day for reinforcement probabilities of .50 (for both stimuli) and performance did not improve anymore.

The reinforcement probabilities that were used in Experiment 1 were .10, .50, and .90. Four pairs of asymmetric reinforcement probabilities ("conditions") were presented to each animal, and each condition was in effect for five consecutive days. The sequence of

TABLE 1 Testing sequences in Experiment 1 (uncontrolled reinforcement ratio) for each of the four animals.

	Testing sequences			
Subject ID	**1**	**2**	**3**	**4**
AD1	.50 / .10	.50 / .90	.90 / .50	.10 / .50
AD2	.50 / .90	.50 / .10	.10 / .50	.90 / .50
AD3	.90 / .50	.10 / .50	.50 / .10	.50 / .90
AD4	.10 / .50	.90 / .50	.50 / .90	.50 / .10

Note: Numbers in cells give reinforcement probabilities for correct S1 and correct S2 trials, respectively.

TABLE 2 Testing sequences in Experiment 2 (controlled reinforcement ratio) for each of the four animals.

	Testing sequences			
	VR 2		**VR 6**	
Subject ID	**1**	**2**	**3**	**4**
AD1	.75 / .25	.25 / .75	.75 / .25	.25 / .75
AD2	.25 / .75	.75 / .25	.25 / .75	.75 / .25
AD3	.75 / .25	.25 / .75	.75 / .25	.25 / .75
AD4	.25 / .75	.75 / .25	.25 / .75	.75 / .25

Note: Numbers in cells specify the relative reinforcement probabilities for correct S1 and correct S2 trials, respectively. All subjects first underwent two conditions with VR 2 and then two conditions with VR 6.

conditions was counterbalanced across animals and is given in Table 1. Before the first experimental condition was run, all animals underwent 3 days of baseline testing with reinforcement probabilities of both stimuli set to .50. At the conclusion of the experiment, another 2 days of baseline testing were conducted.

Experiment 2 was similar to Experiment 1, using the same task with the same auditory stimuli, but reinforcement was provided according to a controlled reinforcer ratio schedule at intervals determined by two different variable ratio (VR) schedules. Generally speaking, a VR *N* schedule of reinforcement specifies that a reinforcer becomes available after *N* responses. Here, we followed previous authors (McCarthy & Davison, 1984) and implemented a reinforcement schedule where, after each reinforcement, the next reinforcement was assigned to either of the two response ports with a certain probability, and then a variable number of correct responses (VR *N*) toward that port had to be emitted to obtain that reinforcement. Until that happened, no reinforcement could be obtained at either port. This procedure ensures that the relative reinforcement ratio for the two ports is fixed throughout the session. We used relative reinforcement ratios of 1:1, 3:1, and 1:3 (corresponding to relative reinforcement probabilities of .50 vs. .50, .75 vs. .25, and .25 vs. .75). Put differently, as long as the animal emits a minimal number of correct responses in a session, its behavior has no influence on the relative reinforcement ratio. This contrasts with the standard uncontrolled reinforcer ratio schedule, where the animal's relative response ratio directly influences the relative reinforcement ratio (e.g., the more the animal emits R1, the more reinforcers it will get for R1 and the less for R2).

We ran asymmetric reinforcement conditions with two different VR schedules. Thus, these conditions differ in reinforcement density (maximum number of reinforcers that can be obtained per trial). For VR 2, the number of required correct responses at a given port until reinforcement could be retrieved ranged from 1 to 9, with a mean of 2. For VR 6, the number of correct responses at a given port until reinforcement could be retrieved ranged from 1 to 43, with a mean of 6.

Experiment 2 began with 3 days of controlled-reinforcement-ratio baseline testing where reinforcers became available according to a VR 2 schedule, and reinforcers were assigned to the two ports with equal probability. Then, the animals underwent two conditions with VR 2 in which reinforcement probabilities were .25 for S1 and .75 for S2 or vice versa. Next, the animals underwent two conditions with the same probabilities but with a lower reinforcement density (VR 6). In each pair of conditions, the order of conditions was counterbalanced across animals and is given in Table 2.

Data analysis and model fitting

All analyses and model fits were conducted in Matlab. The experimental data consisted of series of stimulus presentations, event time stamps, and binary choices. The main indices of behavioral performance were the proportions of correct responses—*P*(correct), calculated as the unweighted mean of *P*(R1|S1) and *P*(R2|S2)—and the proportions of R2—*P*(R2), calculated as the unweighted mean of *P*(R2|S2) and *P*(R2|S1). The SDT indices were calculated from all completed trials in each session using standard formula:

$$d' = \Phi^{-1}(\mathrm{HR}) - \Phi^{-1}(\mathrm{FAR}),$$

where Φ^{-1} denotes the normal inverse cumulative distribution function; HR denotes the hit rate on S2 trials, *P*(R2|S2); and FAR denotes the false-alarm rate on S1 trials, *P*(R2|S1).

Relatedly, the criterion was calculated as

$$c = -0.5 \times \left(\Phi^{-1}(\mathrm{HR}) + \Phi^{-1}(\mathrm{FAR})\right).$$

Optimal criteria (criteria at which expected reinforcement was maximal given a certain value of d' and a programmed reinforcement probabilities) were calculated by custom-written Matlab code through numerical optimization.

Details of the fitting procedure for the three SDT-based criterion learning models have been described in earlier studies (Stoilova et al., 2020; Stüttgen et al., 2013). Briefly, the criterion learning models for a fixed leak parameter γ can be formulated as a generalized linear model with a unique maximum, and therefore the likelihood can be maximized reliably with standard numerical optimization procedures (Dorfman, 1973). By repeating this procedure for different γ, the overall likelihood for all parameters can be maximized. The goodness of fit of the three criterion learning models was compared through calculation of the Bayesian information criterion (BIC), a dimensionless measure based on the maximum-likelihood estimate and controlling for different numbers of parameters in competitor models:

$$\mathrm{BIC} = 2 \times \mathrm{NLL} + k \times \log(N),$$

where NLL is the negative log likelihood of the data under the model using the best-fitting parameters, k is the number of free parameters (four for Models 1 and 2: γ, the model-specific learning-rate parameter [δ or υ], and the two means of the stimulus distributions and five parameters for Model 3, which features two learning rates), and N is the number of trials. Models with smaller BIC values are preferred. As per convention, the strength of evidence against the model with the higher BIC value is judged to be "unimportant" if difference in BIC is 0–2, "positive" for differences of 2–6, "strong" for differences of 6–10 strong, and "very strong" if the difference exceeds 10 (Kass & Raftery, 1995).

For Experiment 1, we computed model predictions of the steady-state criterion positions. The responses are stochastic, so there is an equilibrium distribution for the criterion. Although deriving this distribution (and showing that it actually exists) is beyond the scope of this article, heuristically the following can be said: In the steady state, the expected criterion update is zero. So, the expected update due to a reinforced response (for Model 1) or non-reinforced response (for Model 2) is balanced out by the term with the leakage factor that pulls the criterion back to its neutral position.

For Model 1 this means

$$E[\gamma \times c + \delta \times (\mathrm{Rf}_{\mathrm{R1}} - \mathrm{Rf}_{\mathrm{R2}}) - c] = 0,$$

and therefore

$$\gamma \times c - c + \delta \times (\mathrm{E}[\mathrm{Rf}_{\mathrm{R1}}] - \mathrm{E}[\mathrm{Rf}_{\mathrm{R2}}]) = 0,$$

and for Model 2,

$$E[\gamma \times c + \upsilon \times (\mathrm{noRf}_{\mathrm{R2}} - \mathrm{noRf}_{\mathrm{R1}}) - c] = 0,$$

and therefore

$$\gamma \times c - c + \upsilon \times (E[\mathrm{noRf}_{\mathrm{R2}}] - E[\mathrm{noRf}_{\mathrm{R1}}]) = 0.$$

Thus, there is a linear relation between the criterion position in the steady state and the difference of the probabilities for (non)reinforcement for a Category 1 response and a Category 2 response:

$$E[\mathrm{Rf}_{\mathrm{R1}}] - E[\mathrm{Rf}_{\mathrm{R2}}] = (1-\gamma)/(\delta \times c),$$

and

$$E[\mathrm{noRf}_{\mathrm{R2}}] - E[\mathrm{noRf}_{\mathrm{R1}}] = (1-\gamma)/(\upsilon \times c)$$

for the two models, respectively.

For Experiment 1, the expected values in these equations are straightforward to determine for a given c. In each trial, the expected probability of obtaining reinforcement when responding correctly is fixed, so

$$E[\mathrm{Rf}_{\mathrm{R1}}] = P(\mathrm{Rf}|\mathrm{S1},\mathrm{R1}) \times P(\mathrm{R1}|\mathrm{S1}) \times P(\mathrm{S1}).$$

The solutions to these steady-state equations were calculated by custom-written Matlab code through numerical optimization. A graphical example is shown in Figure A2.

For Experiment 2, calculating the models' criterion predictions is not easily possible because the expected probability to obtain reinforcement when responding correctly depends on the history of stimuli and responses from previous trials, so there are no trial-independent expressions for $E[\mathrm{Rf}_{\mathrm{R1}}]$, $E[\mathrm{Rf}_{\mathrm{R2}}]$, $E[\mathrm{noRf}_{\mathrm{R1}}]$, and $E[\mathrm{noRf}_{\mathrm{R2}}]$. Therefore, we turned to estimating the expected criterion values through steady-state criterion values that we obtained from forward simulations. To that end, each model encountered each of the two experiments 100 times, and expected criteria were obtained as the averages of the 100 criterion values computed over the last 500 trials of each condition (each condition encompassed 1,500 trials).

Response ratios and reinforcement ratios for assessing the fit of the Davison–Tustin (DT) model (Davison & Tustin, 1978) were taken from the last two sessions of each condition and were computed separately for S1 and S2 trials. A linear fit was determined for each stimulus type, and the DT sensitivity and bias parameters were determined from the fitted values as described by the original authors:

$$\text{In S1 trials}, \log \frac{P(R2|S1)}{P(R1|S1)} = a_1 \times \log \frac{Rf_2}{Rf_1} + \log c - \log d;$$

$$\text{in S2 trials}, \log \frac{P(R2|S2)}{P(R1|S2)} = a_2 \times \log \frac{Rf_2}{Rf_1} + \log c + \log d.$$

Log d represents the vertical separation between the two fitted lines, log c represents the overall (stimulus-independent) response bias (not to be confused with the criterion c in the signal detection models above), and a_1

and a_2 represent the degree to which the response ratio changes when the obtained reinforcement ratio changes. The response probabilities on the left can be estimated from the data. Rf_1 and Rf_2 are the reinforcement rates for the R1 and R2 responses, respectively. Reinforcement rates were also estimated from the data; for the controlled reinforcement ratio schedule, they are by design almost identical to the programmed rates. With these empirical estimates, all the parameters can be found by linear regression (but see Davison & McCarthy, 1981). From the resulting regression lines, we can in turn determine the predicted response probabilities of the fits for each condition. In this case, the log odds of the predicted response probabilities are directly given by the regression line. In the uncontrolled reinforcement ratio case, $Rf_1 = E[Rf_{R1}] = P(Rf,S1,R1) = P(Rf|S1,R1) \times P(S1) \times P(R1|S1)$, and therefore both sides of the regression equation depend on the response probabilities $P(R1|S1)$, and equivalently for Rf_2. Thus, the predicted response probabilities have to be obtained by solving the fixed-point equation, which can easily be done numerically (e.g., by fixed-point iteration or root finding). The resulting response probabilities can then be used to compute predicted criterion values for a signal detection model.

We further implemented a trial-by-trial version of the DT model. In this model, estimates for Rf_1 and Rf_2 in each trial are computed by leaky integration of the reinforcements obtained in past trials—that is, $Rf_1(t) = \gamma \times Rf_1(t-1) + Rf_{R1}(t)$ and $Rf_2(t) = \gamma \times Rf_2(t–1) + Rf_{R2}(t)$, where $Rf_{R1}(t)$ is 1 if a reward was obtained for an R1 response in trial t and 0 otherwise, and analogously for Rf_{R2}. Then, $\log(P(R2|S) / P(R1|S))$ is computed according to the DT model for the stimulus S presented in trial t and a response is sampled randomly with the probability for an R2 response being $P(R2|S)$. For a given γ, the DT model is, again, a generalized linear model with a convex NLL, so the maximum-likelihood fit for the model can be found reliably through numerical optimization.

RESULTS

The left panels in Figure 2 show d' and c for all four subjects and for both experiments (see Figure A1, Panel B, for the proportions of correct trials, R2, and aborted trials). Qualitatively, a few observations are obvious. First, asymmetric reinforcement probabilities induced response biases (c) in all animals, and these were more pronounced in Experiment 1. Second, discrimination performance (d') did not change as prominently and systematically as did response bias, but there seemed to be a trend toward increasing performance over time.

To examine the extent to which d' increased over the course of the experiments, we performed linear regression analysis on all d' values as a function of session number, separately for each animal. The regression coefficients were statistically significant in three of the four animals ($p < .02$ for AD1, AD3, and AD4, and $p = .13$ for AD2) and positive in three out of four animals (0.034, −0.003, 0.018, and 0.008 for rats AD1 through AD4, respectively). Thus, despite several months of previous training, three of the four animals exhibited further increments in d' over the course of testing, although the increase was small within an individual experiment.

The right panels in Figure 2 show ROC plots for each animal, using data from all experimental sessions. We computed ROC curves for the equal-variance Gaussian model and calculated grand d' from averages of $P(R2|S2)$ and $P(R1|S2)$ of all experimental sessions. Overall, grand d' varied from 1.7 (AD2) to 3.7 (AD1), and the distributions of individual sessions' data points were consistent with a singular curvilinear ROC curve. We thus proceeded to apply SDT-based analyses.

Comparison of three trial-by-trial criterion learning models

We compared the three criterion learning models in their ability to fit the experimental data. As detailed in Introduction, Model 1 learns exclusively after all reinforced responses, Model 2 learns exclusively after all nonreinforced responses, and Model 3 learns after both. We here report data from fits to the complete data of each subject (i.e., both Experiment 1 and 2) for convenience, as fitting the models separately for the two experiments produced similar results.

Figure 3 shows the fits of the three models to the data of each animal. Qualitatively, the best fit was achieved by Model 3, which was expected given that it is the only model that has two learning parameters. Model 1 was able to capture the general direction of the bias but the fits diverged markedly in some conditions. Finally, Model 2 fared worst by far.

To compare the model fits quantitatively, we computed the Bayesian information criterion (BIC) for each fit. Figure 4A shows the BIC values for each of the four rats' data fit by each of the three models. Confirming visual inspection, Model 2 was dramatically inferior to Model 1 (BIC differences from 500 to 1,000), which again was clearly inferior to Model 3 (BIC differences of 300 to 500), and this was the case for all animals, although all models produced similar estimates of d' and γ for a given animal (Figure 4B, left and middle panels). Examination of the learning rate parameters revealed that, although the values were always positive for criterion shifts after reinforcement (δ, Models 1 and 3), the values for criterion shifts after reinforcement omission (υ, Models 2 and 3) were, with a single exception, all negative. Importantly, negative learning parameters do not make sense from a theoretical point of view. In the present experimental scenario, they imply that an agent, learning from

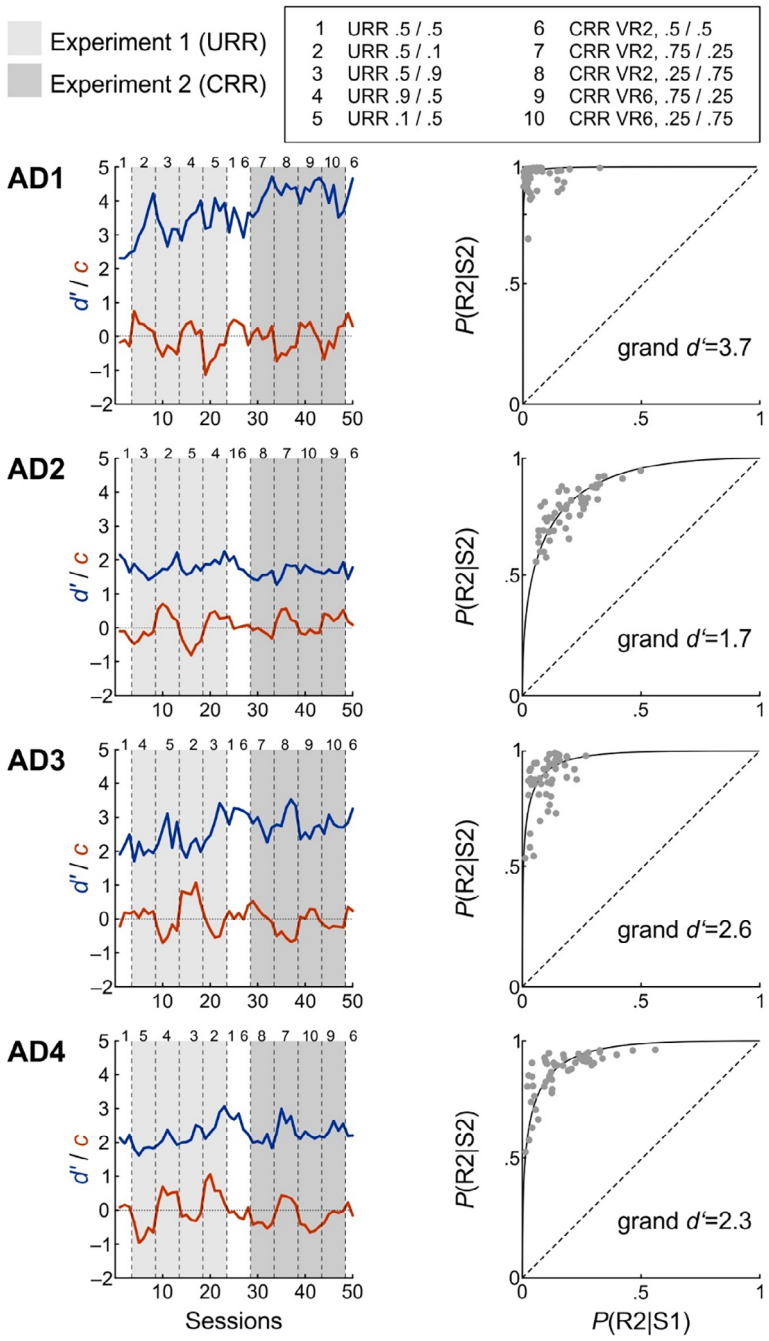

FIGURE 2 Sensitivity (d'), criterion (c), and ROC plots for all subjects for Experiments 1 and 2. Left panels: SDT indices d' and c for subjects AD1 through AD4 for both experiments. Sessions belonging to Experiment 1 and Experiment 2 are highlighted by light gray and dark gray shading, respectively. Conditions within experiments are separated by vertical dashed lines in each panel and are referenced by numbers (see inset). Right panels: ROC plots for the four animals. Each data point represents $P(R2|S2)$ (ordinate) and $P(R2|S1)$ (abscissa) in one of 50 individual sessions. The ROC curves were calculated based on the grand d' for each animal.

nonreinforced responses, shifts the criterion such that the same nonreinforced response is more likely to be produced again in the next trial. In a scenario in which most correct responses are reinforced and incorrect responses are not reinforced, this would increase the likelihood of obtaining no reinforcement in the next trial as well, which we confirmed through forward simulations for our experimental conditions (Figure A3).

Comparison of predicted and observed steady-state criterion values

In the preceding section, we examined how well the criterion learning models fit the behavioral data on a trial-by-trial basis. The fits were rather poor for Model 1 and Model 2 compared with Model 3, but the negative learning rates for υ in Model 3 do not make sense for

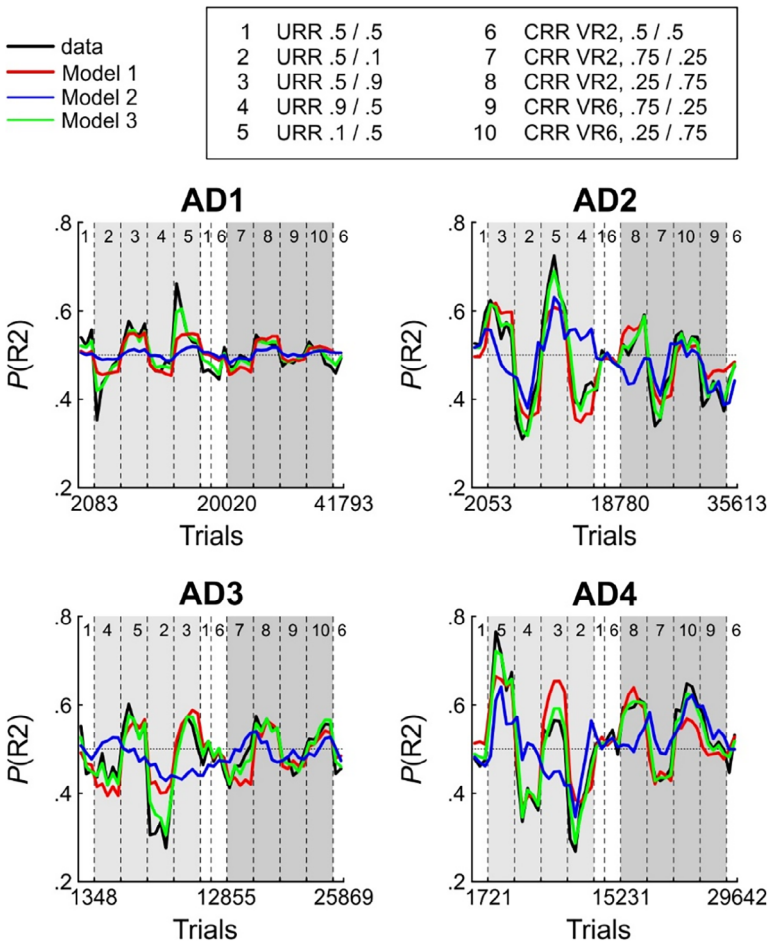

FIGURE 3 Maximum-likelihood fits for the three SDT-based criterion learning models. The models are color coded, and the empirical *P*(R2) values are redrawn in black from Figure 2. All conventions as in Figure 2.

theoretical reasons. This analysis, however, does not tell us why Model 1 fails so badly. Another way to evaluate the models is to ask to what extent their predicted steady-state criterion values align with the empirically obtained steady-state criterion values. As explained in Methods and demonstrated in Figure A2, the models can be used to predict what value the criterion will converge to for a given a set of model parameters (stimulus means, γ, δ, and υ) and reinforcement contingencies. In general, this constitutes an additional and important way to evaluate the models, as a successful model fit does not guarantee that a model endowed with the fitted parameters will generate behavior that is quantitatively and qualitatively similar to that of the subject if confronted with a different sequence of stimuli and/or reinforcements than those used for the fit (see Corrado et al., 2005, for an instructive example). Here, we additionally wanted to test whether the poor fit of Model 1 might be explained by its inability to explain the steady-state behavior of the animals.

Therefore, we generated steady-state predictions for the three criterion learning models for each subject's fitted parameters and all conditions in both experiments. Furthermore, we compared actual criteria with optimal criteria that provide a convenient benchmark for gauging performance. The assumption that over the course of evolution nonoptimal strategies have been eliminated (as they are by definition inferior to optimal ones) may be taken to imply that the brain indeed instantiates an optimal-choice algorithm and therefore represents certain quantities such as posterior odds (Harley, 1981; Yang & Shadlen, 2007). In nonhuman animals, stimulus discrimination performance has been found to come close to optimality in some settings (Feng et al., 2009; Stüttgen, Yildiz, et al., 2011) but not in others (Stoilova et al., 2020; Stüttgen et al., 2013; Teichert & Ferrera, 2010).

Figure 5 shows scatterplots of predicted against actual criterion values for each model. For Models 1 and 3 as well as optimality, the correlations between predicted and actual values were high and statistically significant for all criterion-setting accounts in both experiments (r ranged from .74 to .92, all p values were < .001). In contrast, the predicted and actual criterion models for

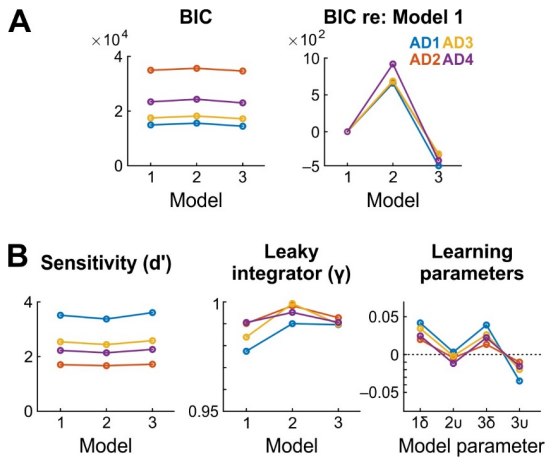

FIGURE 4 Quantitative comparison of the criterion learning models. (A) The left panel shows BIC values for each of model for each animal. The right panel shows the same but normalized to Model 1 for better comparability. The model with the lowest BIC is considered to provide the best fit. (B) Left: sensitivity (d') as estimated from each of the three models for each of the four animals. Middle: best-fitting γ values. Right: best-fitting δ and υ values.

Model 2 were negatively correlated, as expected based on the negative learning rates obtained from the fits of Model 2 and the simulations shown in Figure A3, again demonstrating the inadequacy of this model ($r = -.78$ and $r = -.31$ for Experiment 1 and Experiment 2, respectively). We will not consider Model 2 any further.

A satisfactory model should not only proffer a high positive correlation between predicted and obtained values but also correct parameter estimates, and we can assess the precision of the predictions by the slopes and intercepts of the regression lines. Although all the regression lines had intercepts very close to 0 (ranging from −0.03 to 0.02), only Model 3 produced slopes that were not significantly different from 1 in both experiments (slopes were 0.95 ($p = .79$) and 1.12 ($p = .56$) for Experiments 1 and 2, respectively). The slopes of Model 1 and the optimality account in Experiment 1 were 0.67 and 0.61, and both were significantly smaller than 1 ($p = .004$ and $p = .0004$, respectively). Thus, the bad performance of Model 1 can indeed be attributed to its inability to explain the steady-state behavior of the animals.

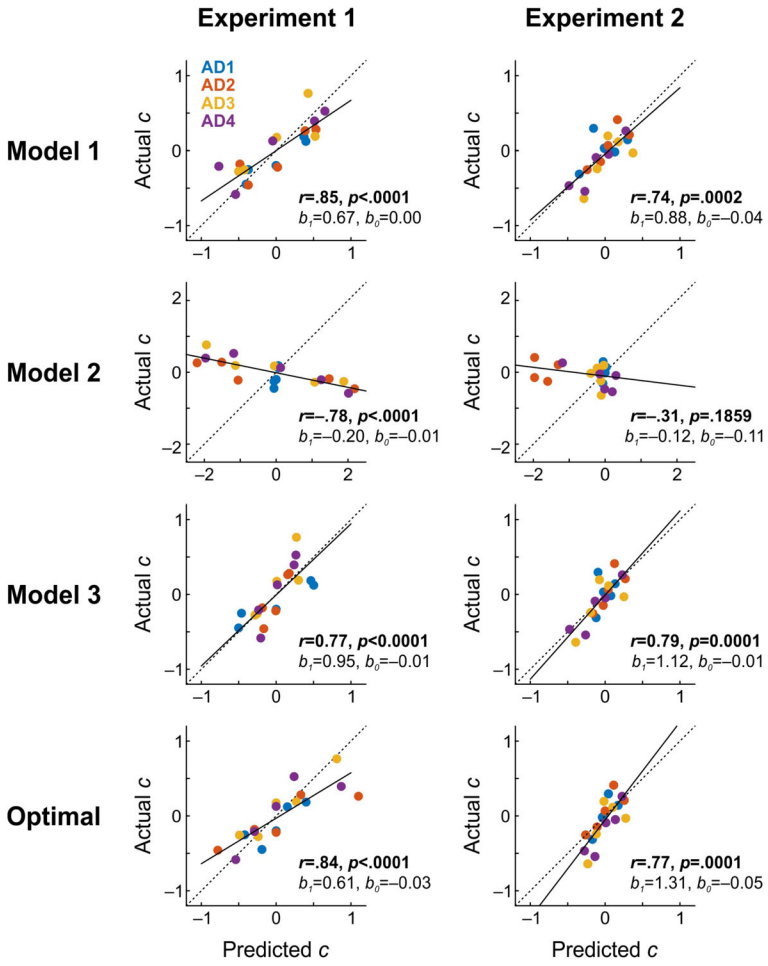

FIGURE 5 Comparison of predicted and actual steady-state decision criteria for the SDT-based models and optimality. Data points represent criterion values calculated over all trials of the last two sessions of each condition. In each panel, the main diagonal (dashed line) represents equality; the black solid line represents the linear regression; and the numbers above panels denote values of the correlation coefficient (r) and its p value, the regression slope (b_1), and its intercept (b_0). Note that the axis ranges for Model 2 had to be expanded and therefore are different from those of all other panels.

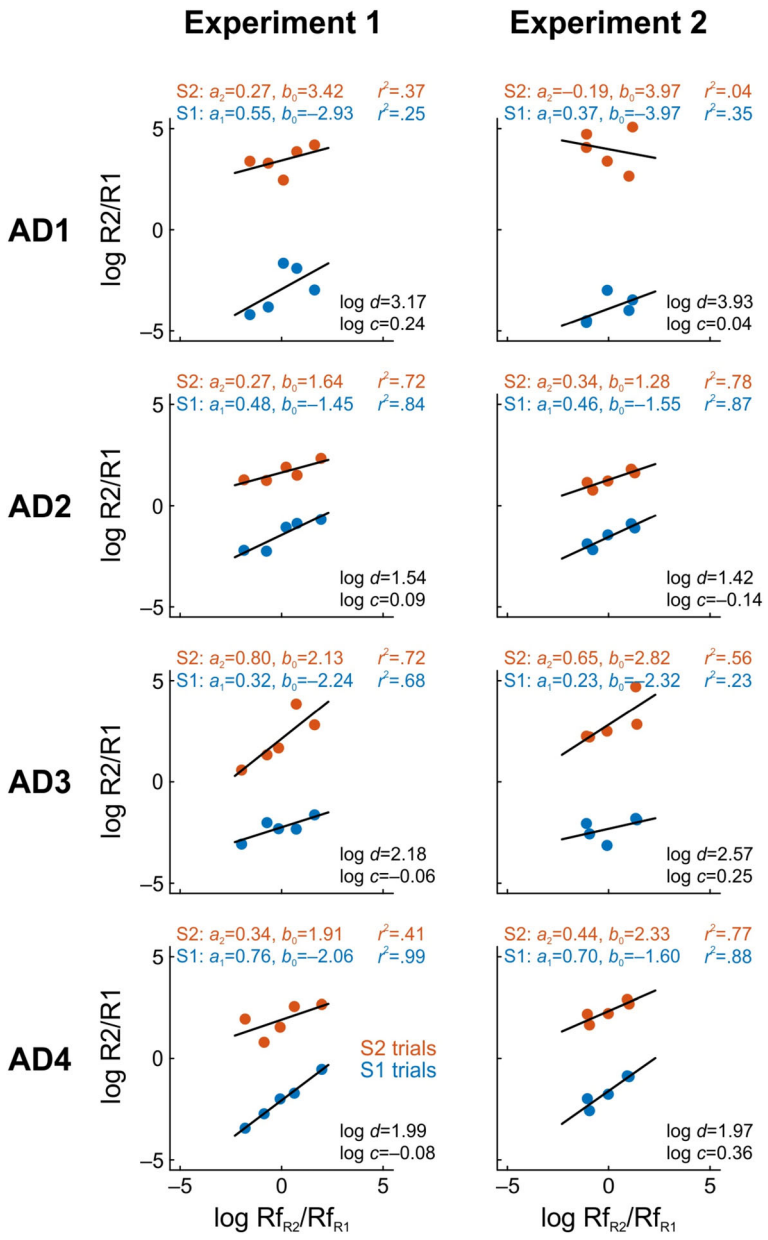

FIGURE 6 Fits of the Davison–Tustin model. In all panels, the data points show the logarithmic reinforcement ratios, $\log(Rf_{R2}/Rf_{R1})$ plotted against the logarithm of the response ratios, log(R2/R1), computed over the total responses in the last two sessions of each condition and separately for S1 (blue) and S2 (red) trials (two lines per panel). The left column panels pertain to Experiment 1 and the right column panels to Experiment 2. Log d and log c are the parameters of the DT model, parameters a_1 and a_2 represent regression slopes in S1 and S2 trials, respectively, b_0 represents the corresponding intercepts, and r^2 is the squared correlation coefficient.

Unfortunately, the close correspondence between predicted and obtained criteria for Model 3 does not automatically validate it. Recall that the fits of this model yielded negative values for the learning-rate parameter υ for all four animals (Figure 4), and as argued above, a negative learning parameter not only violates common sense, but simulating Model 3 with negative υ also fails to generate behavior that is similar to that of the animals (Figure A3). To summarize, none of the three SDT-based criterion learning models provides a satisfactory account of the experimental data.

A trial-by-trial version of the Davison–Tustin model

Although comparatively little research effort has been directed toward describing trial-by-trial changes of response bias in perceptual discrimination tasks, there is an established framework for modeling steady-state response bias as a function of experimental parameters (reviewed in Commons et al., 1991). In a seminal article, Davison and Tustin (1978) proposed a behavioral model of signal detection based on the generalized matching law

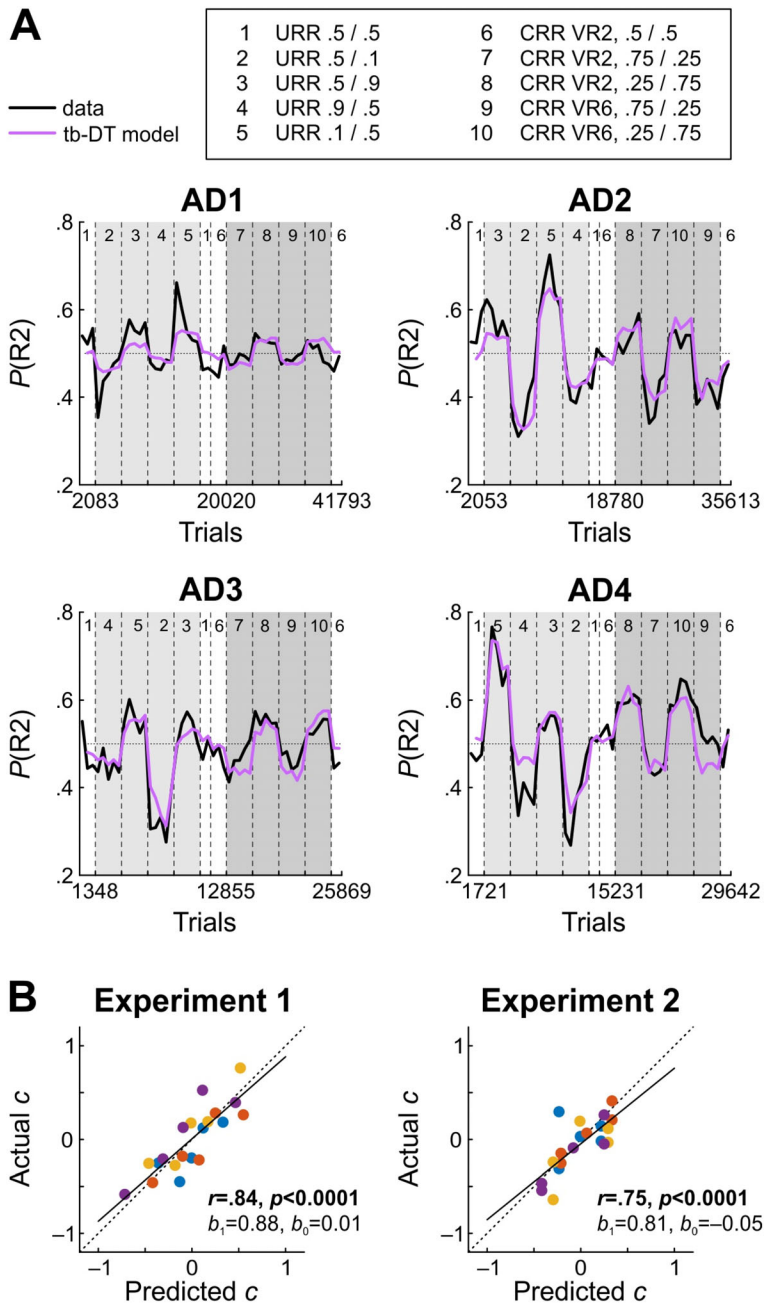

FIGURE 7 Maximum-likelihood fits of the trial-based version of the DT model and steady-state predictions. (A) Fits of the trial-based DT model to each subject's data. Empirical *P*(R2) values redrawn in black from Figure 2. (B) Predicted steady-state criterion values derived based on each subject's fitted tb-DT model parameters. Conventions as in Figure 5.

(Baum, 1974). Their model (henceforth, DT model) predicts that animals match response ratios to reinforcement ratios (as in generalized matching) for each of the stimuli and that the degree of matching is affected by stimulus discriminability (expressed as a sensitivity parameter log *d*), sensitivity to reinforcement (slope *a*), and a general bias parameter (intercept log *c*; see Methods for details). Unlike standard depictions of discrimination behavior, the DT model describes performance in a coordinate system formed by the logarithm of the reinforcement ratio ($\log(Rf_{R2}/Rf_{R1})$) on the abscissa and the logarithm of the response ratio (log(R2/R1)) on the ordinate. The DT model successfully describes behavioral performance in a wide range of different experiments (e.g., Davison & McCarthy, 1980; McCarthy & Davison, 1980, 1984; McCarthy et al., 1982) and has since then been modified and extended to encompass multistimulus, multiresponse procedures (e.g., Alsop, 1991; Davison, 1991; Davison & Jenkins, 1985; Davison & Nevin, 1999). Prompted by the failure of the three SDT-based models, we turned to assess the fit of the DT model to steady-state response bias.

Figure 6 shows each animal's steady-state response ratios (calculated over the last two sessions of each condition) as a function of the obtained reinforcement ratio in logarithmic coordinates for both Experiment 1 (left column) and Experiment 2 (right column). The DT model can be fitted easily through separate linear regressions for S1 and S2 trials. Most of the response data is reasonably well captured by linear fits, with the notable exception of S2 trials in Experiment 2 of subject AD1.

To the best of our knowledge, there is no trial-by-trial instantiation of the DT model so far. Therefore, we developed such a trial-based version of the DT model (henceforth, tb-DT model) to investigate whether this provides a more satisfactory account of the data than do the three SDT-based models. Our tb-DT model uses leaky integration of the history of past reinforcements to estimate the reinforcement ratio in each trial, the same mechanism employed by the SDT-based models (parameter γ). The choice of R1 over R2 in trial t is made probabilistically based on the estimate of the reinforcement ratio (multiplied by a_x, a parameter of sensitivity to reinforcement for stimulus x), the stimulus discriminability parameter log d, and the stimulus-independent bias parameter log c (see Methods for details).

Figure 7A shows the maximum-likelihood fits of the tb-DT model to the data in the same format as in Figure 3 for the three SDT-based models. The parameter estimates for each animal are given in Table 3. Qualitatively, the tb-DT model provided a better fit to the data than did any of the three SDT-based models. The BIC values of the tb-DT model (see Table 4) were considerably smaller than those of Model 2 (range 613–999) and Model 1 for three of four animals (−47, 182, 301, and 80 for subjects AD1 through AD4, respectively). On the other hand, the BIC values were larger than those of Model 3 for three of four animals (−509, −127, 1, and − 311 for AD1 through AD4, respectively). However, although models with low BIC values are usually preferred, a low BIC value by itself is not sufficient to endorse a model. In the present situation, Model 3 is to be rejected nonetheless, in part because forward simulations with negative learning rates generate behavior that is qualitatively very different from that of the subjects (Figure A3), which however is not the case for simulations of the tb-DT model (Figure A4).

Last, we examined the extent to which the steady-state predictions of the DT model are borne out by the data. The scatterplots in Figure 7B show close correspondence between the two not only in terms of high and significant values of r but also in terms of the slopes of the regression lines, which were 0.88 and 0.81 for Experiments 1 and 2, respectively, and both slopes were not significantly different from 1 ($p = .37$ and $p = .27$).

TABLE 3 Fitted parameters of the trial-by-trial version of the DT model.

	Parameters				
Subject ID	**log *d***	**log *c***	**a_1**	**a_2**	**γ**
AD1	3.186	0.01	0.542	0.378	0.983
AD2	1.43	−0.111	0.487	0.389	0.99
AD3	2.233	0.018	0.35	0.697	0.988
AD4	1.897	0.15	0.641	0.49	0.99

Note: The values of the leaky integration parameter γ can be expressed as exponential half-times in trials, log(0.50) / log(γ). For subjects AD1 through AD4, these correspond to 41, 70, 57, and 72 trials.

DISCUSSION

Although SDT is considered by many to be perhaps "the most towering achievement of basic psychological research of the last half century" (Estes, 2002, p. 15), relatively little work has been directed toward uncovering the mechanisms underlying criterion learning and the attempts conducted so far have yielded incoherent results (compare, e.g., Dorfman & Biderman, 1971; Dorfman et al., 1975; Hautus et al., 2022; Kac, 1962; Stüttgen, Yildiz, et al., 2011; Stüttgen et al., 2013; this study). In an effort to begin clarifying this issue, we chose to compare three simple models of criterion learning with respect to their ability to fit experimental data. Because none of the three models provided a satisfactory fit to the data, we additionally derived a dynamic version of the behavioral detection model (Davison & Tustin, 1978).

TABLE 4 Bayesian information criterion (BIC) for all models.

	Parameters			
Subject ID	**Model 1**	**Model 2**	**Model 3**	**Trial-based DT model**
AD1	14,920	15,580	14,458	14,967
AD2	34,941	35,622	34,632	34,759
AD3	17,488	18,176	17,188	17,187
AD4	23,387	24,306	22,996	23,307

Note: For the fit of the DT model to the data, the first 50 trials were used to estimate the reinforcement ratio for trial 51 and are therefore not included in the fit. The BIC values for the tb-DT model were therefore corrected for the smaller number of trials (N_{Trials}) used: $\text{BIC}_{\text{corr}} = \text{BIC} / (N_{\text{Trials}} - 50) \times N_{\text{Trials}}$.

Do any of the three criterion learning models proffer a satisfactory description of criterion learning?

The income-based Model 1, which updates the criterion on reinforced trials only, provided a reasonable fit, although the fitted choice probabilities differed from the data considerably, especially in Conditions 2 and 5 (Figure 3). Actually, steady-state-observed criterion values were often considerably smaller than predicted ones (Figure 5). Forward simulations showed that this model predicts more extreme criteria in Conditions 3 and 4 than in Conditions 2 and 5, respectively (Figure A3), which was however observed in only two out of four animals. For Experiment 2, the forward simulations predict that Conditions 7 and 8 (with reinforcement delivered at VR 2) produce more extreme criterion values than do Conditions 9 and 10 (VR 6), which implies that reinforcement density (i.e., number of expected reinforcers per trial) affects criterion placement. However, this prediction was again met in only two animals.

On the other hand, fits of Model 2, which updates the criterion after nonreinforced responses only, largely failed to account for the data (Figure 3). Forward simulations of this model demonstrated its inability to produce systematic criterion shifts when supplied with a negative learning rate parameter (Figure A3), and its steady-state predictions aligned badly with the observed criteria (Figure 5).

Model 2 learns after all nonreinforced responses, which in our experiments include a large number of correct responses. Accordingly, one might ask whether animals might learn on error trials only. Learning from errors is considered to be a good description of human criterion-setting performance (Dorfman, 1969; Dorfman et al., 1975; Friedman et al., 1968; Kac, 1962; Killeen et al., 2018; Thomas, 1975). However, in our experiments, an agent purely learning from errors would not be affected by asymmetric reinforcement probabilities *at all* and would therefore produce no shifts in response to changing reinforcement contingencies (see Figure A3 for forward simulations), which is at odds with the systematic biases that we observed. Moreover, fitting a pure error-learning model to the data produced negative learning rates for all four animals (data not shown).

Model 3, learning after both reinforced and nonreinforced responses, had by far the lowest BIC values and provided an excellent fit to the data of all animals, often capturing even minor session-to-session fluctuations (Figure 3). Unfortunately, this model still cannot be considered further because the fitted learning-rate parameter υ turned out negative for all animals. This has the effect that, after an error, the criterion is shifted such that the same error becomes more likely to occur on the next trial. This not only makes no sense from a theoretical point of view; in addition, Model 3 with negative υ badly failed to recapture the observed criterion-setting behavior when simulated forward (see Figure A3; Corrado et al., 2005, for a similar case). The reason that Model 3 fits the data so much better than its competitor, Model 1, is likely because it exploits lose–stay patterns in the experimental data, which arise because of response bias.

In conclusion, none of the three criterion learning models provided a satisfactory account of the behavioral data. Although Model 1 was able to capture and generate criterion shifts in the predicted direction, there is a lot of variance left unaccounted for by this model. Models 2 and 3, on the other hand, could be clearly rejected as viable mechanistic accounts of criterion learning.

Does the DT model proffer an adequate description of criterion learning?

In animal psychophysics, subjects usually work for food or water reinforcers. Thus, from the perspective of the animal, there is no fundamental difference between a perceptual decision-making task and other reinforcement-based choice procedures. Accordingly, it is straightforward to connect SDT to the most important description of reinforcement-based choice behavior, the (generalized) matching law (Baum, 1974; Herrnstein, 1961). This was done in classic work by Davison and Tustin (1978) and then developed further (Alsop, 1991; Davison, 1991; Davison & Jenkins, 1985; Davison & Nevin, 1999).

The DT model has been successful in fitting data from a diverse set of experiments (Davison & McCarthy, 1980; McCarthy & Davison, 1979, 1980, 1981, 1984), including our data (Figure 6). Moreover, it predicts the steady-state criterion values better than do any of the three criterion learning models (compare Figure 5 and Figure 7A), with the exception of Model 3, which is untenable for other reasons. One major difference between the criterion learning models and the DT model is how reinforcers are assumed to determine response bias. The learning models conceptualize criterion placement as resulting from the *difference* in the number of reinforcers obtained from the two responses (scaled by reinforcement density, i.e., the overall frequency of reinforcement, because the criterion drifts back to 0 on trials without reinforcement). In contrast, the DT model posits that response bias results from the *ratio* of reinforcers obtained from the two responses. Incidentally, the DT model does not assume any effect of reinforcement density. Although we did not explicitly test whether reinforcement density has an effect, the highly similar steady-state criterion values for VR 2 and VR 6 in Experiment 2 are consistent with this prediction.

To our knowledge, this is the first study to propose and examine a trial-based version of the DT model. Although the original model has been modified and extended considerably (Alsop, 1991; Davison, 1991; Davison & Jenkins, 1985; Davison & Nevin, 1999), we chose to build on the classic DT model for several

reasons. First, the DT model is considerably easier to fit than are its successor models, and the same holds for the trial-based version. Second, both for our as for other two-stimulus, two-response data sets, the DT model already provides a very good fit (Figures 6 and 7A; e.g., McCarthy & Davison, 1979, 1980). Third, simulations of the latest version of the model (Davison & Nevin, 1999) yielded results that were very similar to those shown in Figure A4.

The tb-DT model provided a better fit for three of the four subjects than did the SDT-based criterion learning models. However, it clearly failed to capture the behavior of subject AD1 in Experiment 1. Conspicuously, this subject had by far the highest values of d' and log d and exhibited prominent and transient extreme values of P (R2) in Conditions 2 and 5, which featured the most extreme reinforcer ratios. Similar "overshooting" behavior was seen in an earlier study with pigeons (Stüttgen, Yildiz, et al., 2011). Also, simulations of the tb-DT model with parameters similar to those of our subjects were qualitatively different from observed behavior (Figure A4); in particular, adaptation to novel conditions was considerably more sluggish, and this did not change much when the simulations were run with lower values of γ.

At a more general level, the DT model has several important limitations that also apply to its trial-by-trial instantiation. First, the sensitivity-to-reinforcement parameter a lacks theoretical justification and has repeatedly been found not to be independent from the discriminability parameter log d (e.g., Davison & Jenkins, 1985). Second, it does not explain how subjects come to be biased by the stimuli in the first place—that is, how they learn to associate S1 with R1 and S2 with R2 (this is much clearer in successor models; see, e.g., Davison & Nevin, 1999). Third, from the point of view of signal detection models, the animals do not know whether they are in an S1 or S2 trial, but the tb-DT model equations use this information to compute the animals' response probabilities. Fourth, the DT model is limited to two-stimulus, two-response procedures (although successors to the DT model are not; see Davison, 1991) and our trial-by-trial instantiation is further restricted to discrete-trial procedures. Fifth, it is not clear how the scope of the DT model could be extended to encompass lapses (Pisupati et al., 2021), reaction times (Hernández-Navarro et al., 2021), or decision confidence (Kepecs et al., 2008; Lak, Hueske, et al., 2020). In the future, some of these limitations may be overcome by equipping an SDT-based criterion learning model with an updating mechanism based on the DT model or its successors.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ETHICS APPROVAL

All procedures were approved by local authorities (Landesuntersuchungsamt Rheinland-Pfalz) and conducted in agreement with German law as well as directive 2010/63/EU of the European Parliament.

ORCID

Maik C. Stüttgen https://orcid.org/0000-0002-7031-262X
Vanya V. Stoilova Eckert https://orcid.org/0000-0003-4823-0465
Luis de la Cuesta-Ferrer https://orcid.org/0009-0004-9692-3615
Christina Koß https://orcid.org/0000-0002-1694-5653
Frank Jäkel https://orcid.org/0000-0002-1355-7663

REFERENCES

Alsop, B. (1991). Behavioral models of signal detection and detection models of choice. In M. L. Commons, J. A. Nevin, & M. C. Davison (Eds.), *Signal detection: Mechanisms, models, and applications* (pp. 39–55). Erlbaum. https://doi.org/10.4324/9780203772430

Alsop, B. (1998). Receiver operating characteristics from nonhuman animals: Some implications and directions for research with humans. *Psychonomic Bulletin & Review*, *5*(2), 239–252. https://doi.org/10.3758/BF03212946

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242. https://doi.org/10.1901/jeab.1974.22-231

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115. https://doi.org/10.1037/a0014351

Boneau, C. A., & Cole, J. L. (1967). Decision theory, the pigeon, and the psychophysical function. *Psychological Review*, *74*(2), 123–135. https://doi.org/10.1037/h0024287

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121*(2), 177–194. https://doi.org/10.1037/0096-3445.121.2.177

Commons, M. L., Nevin, J. A., & Davison, M. C. (Eds.). (1991). *Signal detection: Mechanisms, models, and applications*. Erlbaum. https://doi.org/10.4324/9780203772430

Corrado, G. S., Sugrue, L. P., Sebastian Seung, H., & Newsome, W. T. (2005). Linear-nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior*, *84*(3), 581–617. https://doi.org/10.1901/jeab.2005.23-05

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula Publishing.

Davison, M. (1991). Stimulus discriminability, contingency discriminability, and complex stimulus control. In M. L. Commons, J. A. Nevin, & M. C. Davison (Eds.), *Signal detection: Mechanisms, models, and applications* (pp. 57–78). Erlbaum. https://doi.org/10.4324/9780203772430

Davison, M., & Jenkins, P. E. (1985). Stimulus discriminability, contingency discriminability, and schedule performance. *Animal Learning & Behavior*, *13*(1), 77–84. https://doi.org/10.3758/BF03213368

Davison, M., & McCarthy, D. (1980). Reinforcement for errors in a signal-detection procedure. *Journal of the Experimental Analysis of Behavior*, *34*(1), 35–47. https://doi.org/10.1901/jeab.1980.34-35

Davison, M., & McCarthy, D. (1981). Undermatching and structural relations. *Behaviour Analysis Letters*, *1*(1), 67–72.

Davison, M., & Nevin, J. (1999). Stimuli, reinforcers, and behavior: An integration. *Journal of the Experimental Analysis of Behavior*, *71*(3), 439–482. https://doi.org/10.1901/jeab.1999.71-439

Davison, M., & Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the*

Experimental Analysis of Behavior, *29*(2), 331–336. https://doi.org/10.1901/jeab.1978.29-331

Dorfman, D. D. (1969). Probability matching in signal detection. *Psychonomic Science*, *17*(2), 103–103. https://doi.org/10.3758/BF03336468

Dorfman, D. D. (1973). Likelihood function of additive learning models: Sufficient conditions for strict log-concavity and uniqueness of maximum. *Journal of Mathematical Psychology*, *10*(1), 73–85. https://doi.org/10.1016/0022-2496(73)90005-9

Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology*, *8*(2), 264–284. https://doi.org/10.1016/0022-2496(71)90017-4

Dorfman, D. D., Saslow, C. F., & Simpson, J. C. (1975). Learning models for a continuum of sensory states reexamined. *Journal of Mathematical Psychology*, *12*(2), 178–211. https://doi.org/10.1016/0022-2496(75)90056-5

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review*, *105*(2), 280–298. https://doi.org/10.1037/0033-295X.105.2.280

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin and Review*, *9*(1), 3–25. https://doi.org/10.3758/BF03196254

Feng, S., Holmes, P., Rorie, A., & Newsome, W. T. (2009). Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS Computational Biology*, *5*(2), Article e1000284. https://doi.org/10.1371/journal.pcbi.1000284

Friedman, M. P., Carterette, E. C., Nakatani, L., & Ahumada, A. (1968). Comparisons of some learning models for response bias in signal detection. *Perception & Psychophysics*, *3*, 5–11. https://doi.org/10.3758/BF03212703

Funamizu, A. (2021). Integration of sensory evidence and reward expectation in mouse perceptual decision-making task with various sensory uncertainties. *IScience*, *24*, Article 102826. https://doi.org/10.1016/j.isci.2021.102826

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Harley, C. B. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology*, *89*(4), 611–633. https://doi.org/10.1016/0022-5193(81)90032-1

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2022). *Detection theory: A user's guide* (3rd ed.). Routledge. https://doi.org/10.4324/9781003203636

Hernández-Navarro, L., Hermoso-Mendizabal, A., Duque, D., de la Rocha, J., & Hyafil, A. (2021). Proactive and reactive accumulation-to-bound processes compete during perceptual decisions. *Nature Communications*, *12*(1), Article 7148. https://doi.org/10.1038/s41467-021-27302-8

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267–272. https://doi.org/10.1901/jeab.1961.4-267

Kac, M. (1962). A note on learning signal detection. *IRE Transactions on Information Theory*, *8*(2), 126–128. https://doi.org/10.1109/TIT.1962.1057687

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Killeen, P. R., Taylor, T. J., & Treviño, M. (2018). Subjects adjust criterion on errors in perceptual decision tasks. *Psychological Review*, *125*(1), 117–130. https://doi.org/10.1037/rev0000056

Lak, A., Hueske, E., Hirokawa, J., Masset, P., Ott, T., Urai, A. E., Donner, T. H., Carandini, M., Tonegawa, S., Uchida, N., & Kepecs, A. (2020). Reinforcement biases subsequent perceptual decisions when confidence is low: A widespread behavioral phenomenon. *ELife*, *9*, 1–26. https://doi.org/10.7554/eLife.49834

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, *27*(6), 821–832. https://doi.org/10.1016/j.cub.2017.02.026

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., Reddy, C. B., Kepecs, A., Harris, K. D., & Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron*, *105*(4), 700–711. https://doi.org/10.1016/j.neuron.2019.11.018

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79. https://doi.org/10.1037/h0039723

McCarthy, D., & Davison, M. (1979). Signal probability, reinforcement and signal-detection. *Journal of the Experimental Analysis of Behavior*, *32*(3), 373–386. https://doi.org/10.1901/jeab.1979.32-373

McCarthy, D., & Davison, M. (1980). Independence of sensitivity to relative reinforcement rate and discriminability in signal-detection. *Journal of the Experimental Analysis of Behavior*, *34*(3), 273–284. https://doi.org/10.1901/jeab.1980.34-273

McCarthy, D., & Davison, M. (1981). Towards a behavioral theory of bias in signal detection. *Perception & Psychophysics*, *29*(4), 371–382. https://doi.org/10.3758/bf03207347

McCarthy, D., & Davison, M. (1984). Isobias and alloiobias functions in animal psycophysics. *Journal of Experimental Psychology: Animal Behavior Processes*, *10*(3), 390–409. https://doi.org/10.1037/0097-7403.10.3.390

McCarthy, D., Davison, M., & Jenkins, P. E. (1982). Stimulus discriminability in free-operant and discrete-trial detection procedures. *Journal of the Experimental Analysis of Behavior*, *37*(2), 199–215. https://doi.org/10.1901/jeab.1982.37-199

Mill, R. W., Alves-Pinto, A., & Sumner, C. J. (2014). Decision criterion dynamics in animals performing an auditory detection task. *PLoS ONE*, *9*(12), Article e114076. https://doi.org/10.1371/journal.pone.0114076

Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2021). Lapses in perceptual decisions reflect exploration. *ELife*, *10*, 1–27. https://doi.org/10.7554/ELIFE.55490

Stoilova, V. V., Knauer, B., Berg, S., Rieber, E., Jäkel, F., & Stüttgen, M. C. (2020). Auditory cortex reflects goal-directed movement but is not necessary for behavioral adaptation in sound-cued reward tracking. *Journal of Neurophysiology*, *124*(4), 1056–1071. https://doi.org/10.1152/jn.00736.2019

Stubbs, D. A., & Pliskoff, S. S. (1969). Concurrent responding with fixed relative rate of reinforcement. *Journal of the Experimental Analysis of Behavior*, *12*(6), 887–895. https://doi.org/10.1901/jeab.1969.12-887

Stüttgen, M. C., Kasties, N., Lengersdorf, D., Starosta, S., Güntürkün, O., & Jäkel, F. (2013). Suboptimal criterion setting in a perceptual choice task with asymmetric reinforcement. *Behavioural Processes*, *96*, 59–70. https://doi.org/10.1016/j.beproc.2013.02.014

Stüttgen, M. C., Schwarz, C., & Jäkel, F. (2011). Mapping spikes to sensations. *Frontiers in Neuroscience*, *5*, Article 125.

Stüttgen, M. C., Yildiz, A., & Güntürkün, O. (2011). Adaptive criterion setting in perceptual decision making. *Journal of the Experimental Analysis of Behavior*, *96*(2), 155–176. https://doi.org/10.1901/jeab.2011.96-155

Swets, J. A. (1961a). Detection theory and psychophysics: A review. *Psychometrika*, *26*(1), 49–63. https://doi.org/10.1007/BF02289684

Swets, J. A. (1961b). Is there a sensory threshold? *Science*, *134*(3473), 168–177. https://doi.org/10.1126/science.134.3473.168

Teichert, T., & Ferrera, V. P. (2010). Suboptimal integration of reward magnitude and prior reward likelihood in categorical decisions by monkeys. *Frontiers in Neuroscience*, *4*, Article 186. https://doi.org/10.3389/fnins.2010.00186

Thomas, E. A. C. (1975). Criterion adjustment and probability matching. *Perception & Psychophysics*, *18*, 158–162. https://doi.org/10.3758/BF03204104

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*(1), 68–111. https://doi.org/10.1037//0033-295X.91.1.68

Vandevelde, J. R., Yang, J.-W., Albrecht, S., Lam, H., Kaufmann, P., Luhmann, H. J., & Stüttgen, M. C. (2023). Layer- and cell-type-specific differences in neural activity in mouse barrel cortex during a whisker detection task. *Cerebral Cortex*, *33*, 1361–1382. https://doi.org/10.1093/cercor/bhac141

Wichmann, F. A., & Jäkel, F. (2018). Methods in psychophysics. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., pp. 1–42). John Wiley & Sons. https://doi.org/10.1002/9781119170174.epcn507

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*(7148), 1075–1080. https://doi.org/10.1038/nature05852

How to cite this article: Stüttgen, M. C., Dietl, A., Stoilova Eckert, V. V., de la Cuesta-Ferrer, L., Blanke, J.-H., Koß, C., & Jäkel, F. (2024). Influence of reinforcement and its omission on trial-by-trial changes of response bias in perceptual decision making. *Journal of the Experimental Analysis of Behavior*, *121*(3), 294–313. https://doi.org/10.1002/jeab.908

APPENDIX A

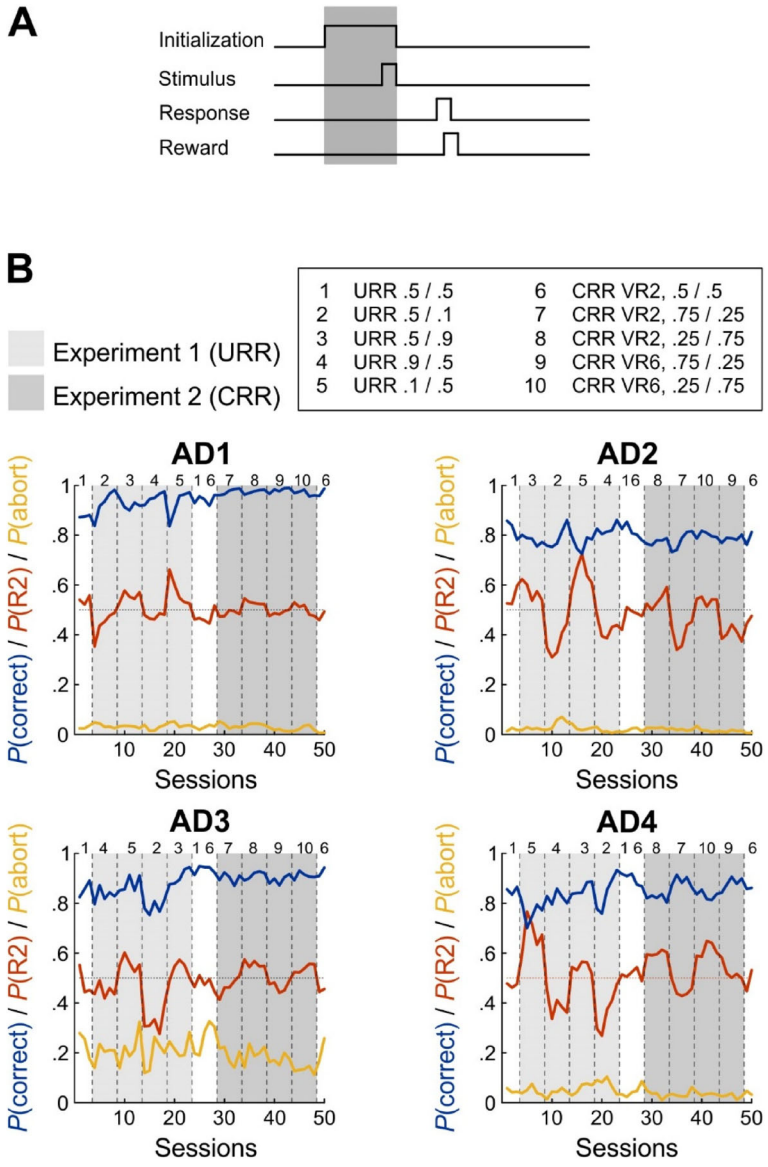

FIGURE A1 Behavioral paradigm and results overview for Experiments 1 and 2. (A) Schematic outline of epochs in a single trial of the behavioral task. Gray shaded area denotes time interval in which rats had to maintain nose poking at the center port. (B) Proportion of correct trials, *P*(correct); proportion of R2, *P*(R2); and proportion of aborted trials, *P*(abort), for all four subjects (AD1 through AD4) for both experiments. All conventions as in Figure 2.

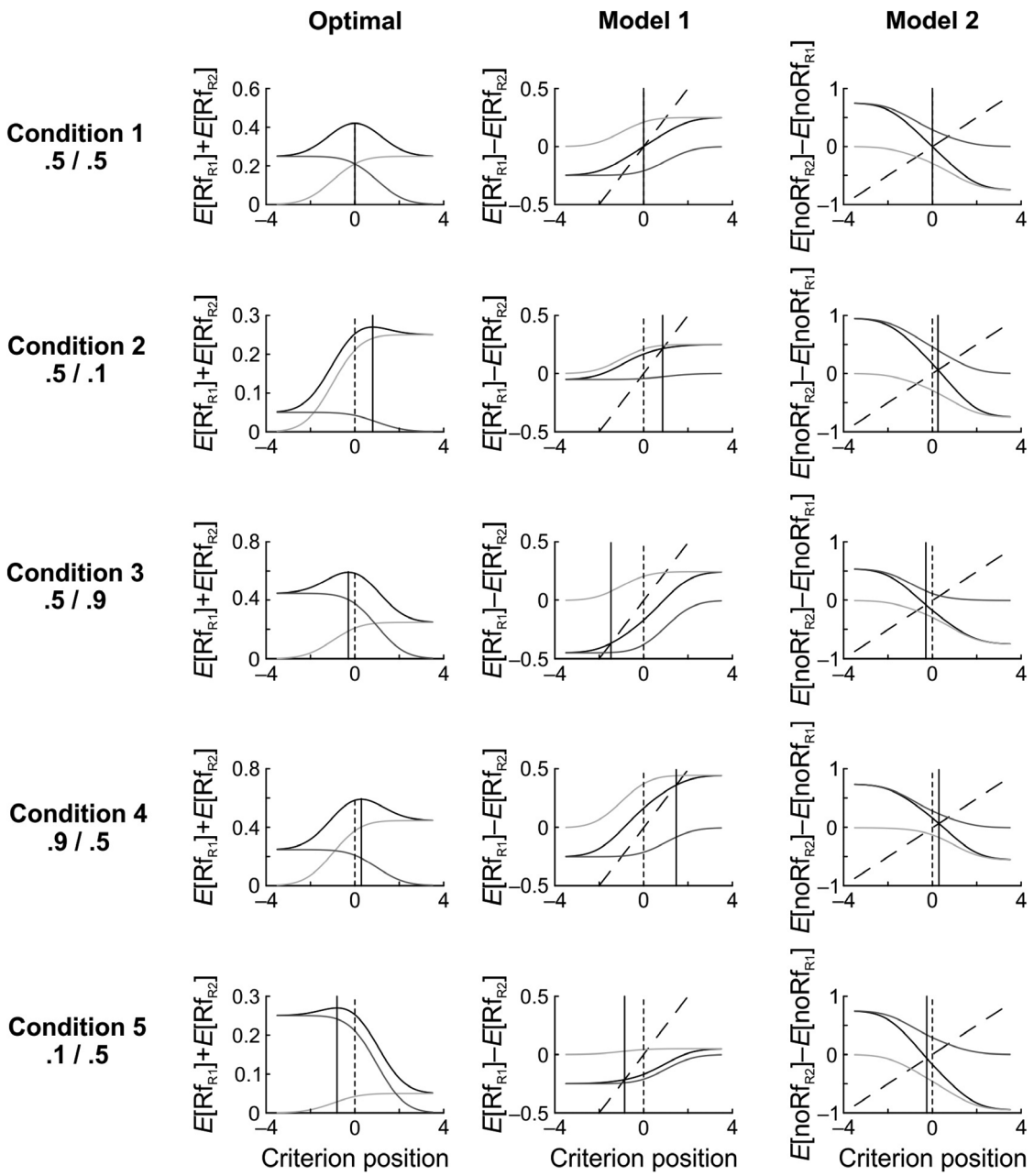

FIGURE A2 Predictions of criterion location for optimal performance as well as Models 1 and 2. In all plots, $d' = 2$ and $\gamma = 0.99$. For Models 1 and 2, $\delta = \upsilon = 0.04$. In the left column, criterion predictions for optimal performance are shown. The bold black line is the objective reward function, which represents the total expected probability of reinforcement in a trial dependent on the criterion. The gray lines are the probabilities for reinforcement following each of the responses, respectively. Optimal performance is achieved at the maximum of the objective reward function; the corresponding criterion is plotted as a thin black line. For reference, the neutral criterion at zero is shown as a dotted black line. In the middle and right column, criterion predictions for Models 1 and 2, respectively, are shown. The difference between the expected probabilities for a trial with/without reinforcement following each response is plotted as a bold black line. The gray lines are these expected probabilities for each response, respectively. Additionally, a straight line through zero is plotted as a dashed black line, whose slope depends on the leakage term γ and the step size δ or υ: it is $(1 - \gamma) / \delta$ for Model 1 and $(1 - \gamma) / \upsilon$ for Model 2. The predicted criterion location for the models is at the intersection of this straight line with the bold black line. For reference, the neutral criterion at zero is shown as a dotted black line. [Corrections made on 22 April 2024, after first online publication: Figure A3 was published as Figure A2. This has been corrected in this version.]

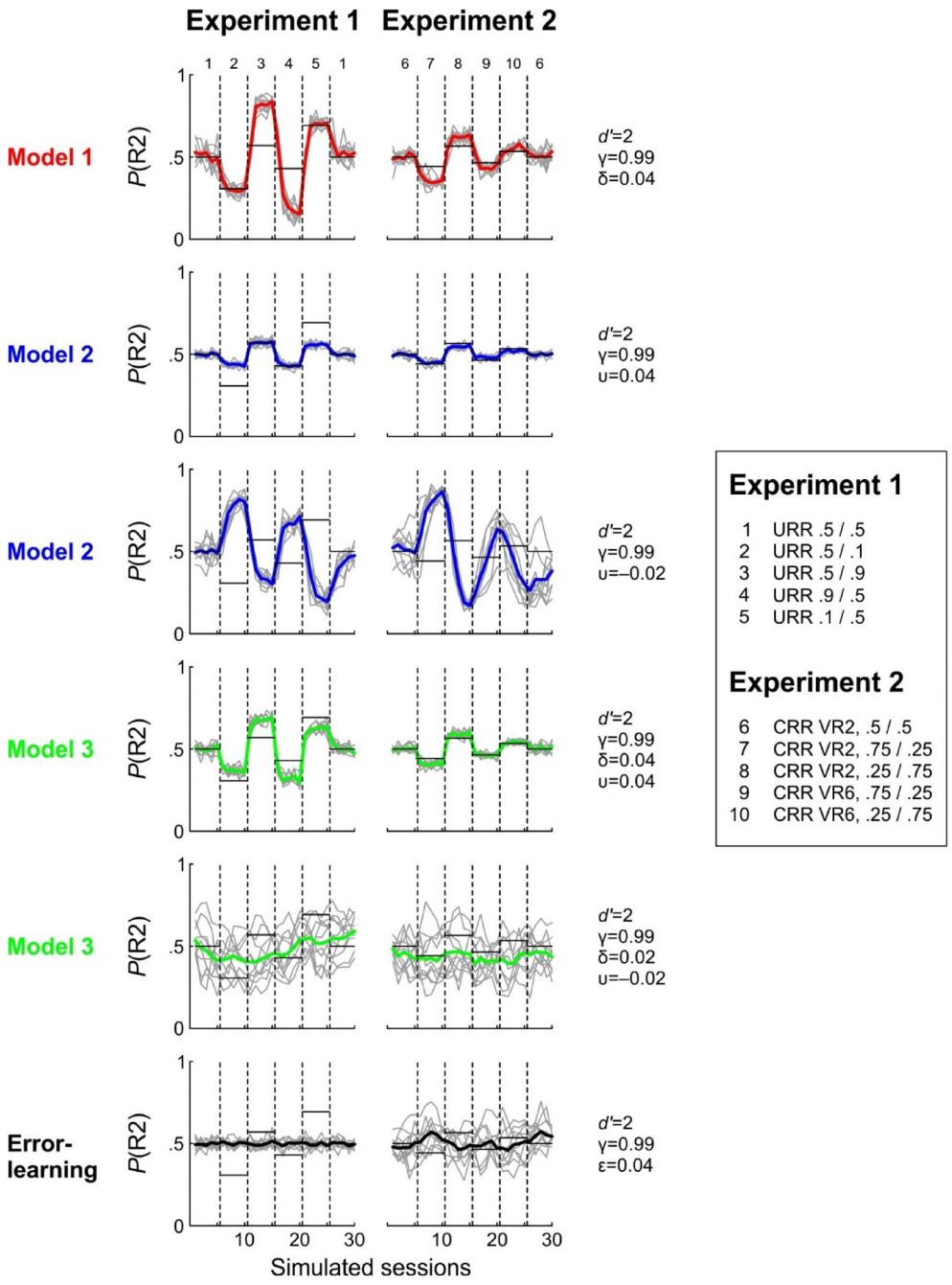

FIGURE A3 Example simulations of Models 1, 2, and 3, and an error-based learning model for both experiments. Each model was simulated forward 10 times with random stimulus sequences. Each condition consisted of 1,500 trials that were split up into blocks of six virtual sessions with 300 trials each. In each panel, the 10 individual simulated sessions are shown as thin gray lines, their averages as bold lines. Dashed vertical lines separate the conditions, and horizontal solid lines give the position of the optimal bias for reference. Model parameters are indicated on the right of each panel. Models 2 and 3 were run twice, once with positive values for both δ and υ and once with a negative value of υ (as was found for the fits to the animals' data). Here, learning-rate parameters δ and υ were reduced to 0.02 and −0.02, respectively, because values of 0.04 and −0.04 consistently produced exclusive choice in all simulations. Additionally, a pure error-learning model (Kac, 1962) was simulated that was identical to Model 2, with the exception that only incorrect rather than all trials without reinforcement led to a criterion shift with magnitude ε. Optimal P(R2) values for Experiment 2 were obtained through numerical optimization over 10^7 trials. [Corrections made on 22 April 2024, after first online publication: Figure A2 was published as Figure A3. This has been corrected in this version.]

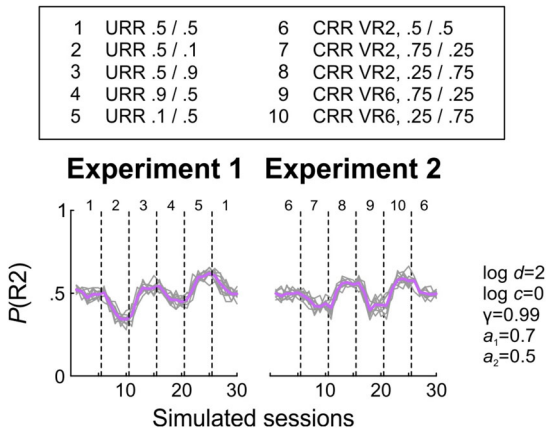

FIGURE A4 Example simulations of the trial-based version of the DT model for both experiments. Model simulations were conducted as described in Figure A2, and the same conventions apply in this figure.