"Play it Again": a new method for testing metacognition in animals

Allison L. Foote and **Jonathon D. Crystal**
University of Georgia, Athens, GA, USA

Abstract

Putative metacognition data in animals may be explained by non-metacognition models (e.g., stimulus generalization). The primary objective of the present study was to develop a new method for testing metacognition in animals that may yield data that can be explained by meta-cognition but not by non-metacognition models. Next, we used the new method with rats. Rats were first presented with a brief noise duration which they would subsequently classify as short or long. Rats were sometimes forced to take an immediate duration test, forced to repeat the same duration, or had the choice to take the test or repeat the duration. Metacognition, but not an alternative non-meta-cognition model, predicts that accuracy on difficult durations is higher when subjects are forced to repeat the stimulus compared to trials in which the subject chose to repeat the stimulus, a pattern observed in our data. Simulation of a non-metacognition model suggests that this part of the data from rats is consistent with metacognition, but other aspects of the data are not consistent with metacognition. The current results call into question previous findings suggesting that rats have metacognitive abilities. Although a mixed pattern of data does not support metacognition in rats, we believe the introduction of the method may be valuable for testing with other species to help evaluate the comparative case for metacognition.

Keywords

Metacognition; Metacognitive control; Simulations; Rats

Introduction

Metacognition has been defined as the ability to reflect upon one's own internal cognitive states (Metcalfe and Kober 2005; Smith 2009). Sometimes people know that they do not know some information. In such cases, the person may avoid some behavior (e.g., do not try to assemble a complicated project) or may seek out additional information (e.g., read the user's manual) before proceeding. Studies of metacognition in people can draw on the participant's self-reported knowledge of his/her own cognitive state or the individual's behavior. By contrast, attempts to study metacognition in nonhuman animals (henceforth animals) can only rely on behavioral observations. Our ability to attribute observed behaviors to metacognition (rather than various alternative explanations) represents a significant inferential challenge (e.g., Hampton 2009).

In this article, we briefly outline an ongoing debate about the types of experiments with animals that can be explained by metacognition and non-metacognition hypotheses. The primary focus of this work is the development of a new experimental approach that may yield data that can be explained by metacognition but not by a non-metacognition framework. Next, we used the new method with rats. Some of the data from rats are consistent with metacognition (suggested by simulations with a non-metacognition model), but other aspects of the data are not consistent with metacognition. Although the pattern of data from rats in the new method is mixed, we believe the introduction of the method may be valuable for testing with other species to help evaluate the comparative case for metacognition.

Significant progress has been made in developing experimental tests of metacognition in animals (e.g., Basile et al. 2009; Beran et al. 2009; Call and Carpenter 2001; Call 2010; Hampton 2001; Hampton et al. 2004; Kornell et al. 2007; Smith et al. 2006, 2010; Washburn et al. 2010). Most of this work has been conducted with monkeys. However, some attempts have been made to document metacognition in other animals (e.g., rats: Foote and Crystal 2007; pigeons: Inman and Shettleworth 1999; Roberts et al. 2009; Sole et al. 2003; Sutton and Shettleworth 2008; dogs: McMahon et al. 2010). Perhaps some of the strongest criticism about experimental techniques and inferential pitfalls was occasioned by the attempt to document metacognition in rats (cf. Smith et al. 2008; see also Crystal and Foote 2009; Hampton 2009). The focus of this article is to develop a method for documenting metacognition that may address this criticism.

Foote and Crystal (2007) sought to determine if rats would decline problems because they knew that they did not know the correct answer. The procedure consisted of a study phase, a choice phase, and a test/small reward phase. Eight logarithmically spaced, white-noise durations ranging from 2 to 8 s were used as study items. During the study phase, rats were presented with a white-noise stimulus duration. Rats could later classify the stimulus duration as either short or long, in order to receive a large food reward, which depended on the match between the recently presented duration and the classification response. In the choice phase, the rat sometimes had the option to choose between taking and declining the memory test. If the rat opted to take a memory test, it interrupted a photobeam in a nose-poke aperture (designated as the take-test response), which resulted in commencement of the test phase. During the test phase, two levers were inserted into the box, and the rat could classify the stimulus duration as short or long by pressing one lever for short or the other lever for long. If the rat was correct it received a large reward of six food pellets; no pellets were delivered if an incorrect duration classification occurred. On the other hand, if the rat selected the option to decline a memory test, by interrupting a photobeam in the other nose-poke aperture (designated as the decline-test response), it proceeded directly to the small reward phase. Instead of taking a memory test during the test phase, the rat was given a less desirable, but guaranteed, reward of one food pellet. On some trials, both take-test and decline-test responses were available, which was signaled by illumination of both nose-poke apertures (and both nose-poke photobeams were active as described above); we refer to these occasions as choice trials. However, on other trials, only the take-test response was available, which was signaled by illumination of the forced-test nose-poke aperture (and the decline-test photobeam was inactive). Foote and Crystal found that rats declined difficult tests more often than easy tests and that accuracy was worse on difficult tests when rats did not have the option to decline. We argued that this pattern of behavior was consistent with metacognition.

Importantly, Smith et al. (2008) subsequently constructed a model (referred to hereafter as the response-strength model) that showed that a non-metacognition model could produce the same patterns of behavior that we observed in rats. Clearly, putative evidence for

metacognition in rats is critically undermined when a non-metacognition model can produce the observed pattern of data. The objective of the current work was to develop a new method for testing meta-cognition in animals that makes a unique prediction (i.e., a prediction not made by a non-metacognition model). Before introducing our new method, we briefly review the essential characteristics of the response-strength model.

The response-strength model uses associative, habit-formation, and generalization principles to predict which response is selected under varying experimental conditions. For the Foote and Crystal (2007) procedure, Smith et al. (2008) proposed that generalization decrements for response strength were modeled by exponential functions. One function declines as durations increase from the anchor stimulus of 2 s for the short response, and another exponential function declines as durations decrease from the other anchor stimulus of 8 s for the long response (see Fig. 1). This is a standard generalization account of the classification task (Shepard 1961, 1987; White 2002), and it produces high levels of accuracy for durations near the anchor stimuli and poorer levels of performance as the durations depart from the anchors. The most difficult problems occur at the subjective middle, which for duration discrimination is generally at the geometric mean of short and long durations (i.e., 4 s; Church and Deluty 1977; Gibbon 1981); the exponent parameters determine the location of the point of subjective equality in the response-strength model. The response-strength model included a low level, flat threshold for the decline response. In contrast to the response strengths for short and long response alternatives, the threshold for the decline response has a constant low level of attractiveness independent of stimulus conditions. The response-strength model uses a winner-take-all response rule. Thus, when two or three response options are available, the model selects the option with the highest response strength. The shape of the functions (and thus the likelihood of any particular response winning in a given stimulus condition) are determined by the exponential parameters and the height of the threshold. Variability is introduced by random variation in the translation of a physical duration into a subjective duration and by random variation in the parameters; these sources of variability may be modeled by sampling from a normal distribution with a specified mean and standard deviation.

Applying the response-strength model to Foote and Crystal's (2007) procedure, we will consider three cases: a very short duration, a very long duration, and an intermediate duration. When the duration is very short (e.g., near 2 s), the response strength for short is very high compared to response strengths for the long and decline thresholds; the model frequently produces a short classification in this case. When the duration is very long (e.g., near 8 s), the response strength for long is very high compared to response strengths for the short and decline thresholds; the model frequently produces a long classification in this case. When the duration is intermediate (e.g., near 4 or 5 s), both short and long response strengths are very low; in this case, the decline threshold will often exceed the response strengths for short and long alternatives, and the model frequently produces a decline response. Smith et al. (2008) simulations showed that decline responses increase as the durations move toward the intermediate range (i.e., as the duration problems become more difficult). Moreover, the simulations showed that accuracy in forced and choice conditions diverged as durations move toward the intermediate range; accuracy was higher for choice trials compared to forced trials, particularly at difficult problems. Note that this pattern of data is predicted by a model that was offered as a non-metacognition proposal. Thus, the response-strength model can explain the putative metacognition data described by Foote and Crystal (2007).

We developed a new method for testing metacognition in animals in close consultation with the response-strength model. The objective was to identify a procedure that could, in principle, generate metacognition data that could not be explained by the response-strength

model. The remainder of this article is divided into two sections. In the next section, we describe our experimental approach and data from rats. In the following section, we describe simulations to compare data from rats in the current experiment with predictions from the response-strength model.

Experiment

Rats were initially presented with a brief noise duration that they would later classify as short or long in a subsequent duration test (Fig. 2). On some trials, rats were forced to immediately take a duration test (referred to as a forced-test trial). On other trials, rats were required to repeat the same brief noise duration that they had just heard (referred to as a forced-repeat trial) before being required to take the test. On a third type of trial, rats had the option to take a duration test immediately or repeat the noise duration (referred to as a choice-take response and choice-repeat response, respectively). One of the critical comparisons in this experiment is accuracy on duration tests that occur after the rats were forced to repeat the stimulus versus accuracy on duration tests that occur after the rats chose to repeat the stimulus; it is in this comparison alone that metacognition and the response-strength model make different predictions, as outlined below.

We assume that animals are sometimes in a low internal state of performance (i.e., the information needed to answer a question is absent) and sometimes in a high internal state of performance (i.e., the information needed to answer a question is present). Metacognition is the hypothesis that animals have access to information about the strength of their knowledge or memory for recently presented information (e.g., in a low state the animal knows that it does not know the answer to the question). We assume that animals choose to repeat the stimulus if they are in a low state of performance on difficult trials. Although a second presentation of a noise duration is expected to increase accuracy, it is assumed that the initial low state of performance also impacts performance at the end of the trial after the rat chooses to repeat the noise duration. Trials in which animals choose to repeat the stimulus function to isolate low states of performance.[1] By contrast, trials in which animals are forced to repeat the stimulus have a combination of low and high states of performance. Therefore, if rats have knowledge about their own cognitive states, they should be less accurate when they have the option to repeat difficult stimulus duration tests than when they are forced to repeat difficult stimulus duration tests. This prediction is unique to difficult stimulus durations because low states of performance rarely occur on easy stimulus durations. In contrast, alternative (i.e., non-metacognitive) proposals predict equal performance on easy and difficult trials when rats have the choice to repeat the stimulus and when they are forced to repeat the stimulus (see "Simulation" section below). We assume that two exposures of the same duration may allow the animal to integrate information from both stimulus presentations and thereby reduce its perceptual error about the stimulus.[2] There are other features of performance that are not uniquely diagnostic of metacognition (i.e., they are also predicted by the response-strength model) that would be needed to provide a clear demonstration of metacognition. For example, metacognition predicts that

---

[1]An animal could maximize the number of pellets obtained by choosing to repeat the stimulus in every trial (thereby obtaining a pellet for selecting the repeat response followed by pellets for a correct duration-classification response). However, this outcome is unlikely to occur because delay discounting (i.e., the observation that reward value declines as a function of delay to reward) in rats (Cardinal et al. 2001; Mazur 1988, 2007; Richards et al. 1997) argues against the choice of a small, immediate reward followed by a large, delayed reward when a large, immediate reward is currently available.

[2]Although we assume that two exposures of the same duration may allow the animal to integrate information from both stimulus presentations and thereby reduce its perceptual error about the stimulus, other possibilities exist. For example, an animal might limit the impact of a second presentation to cases in which it requested a repeat of the stimulus. In our simulations of a response-strength model, we parametrically explore how much weight the animal assigns to first and second stimulus presentations. We consider the full range from all weight assigned to the first stimulus to all weight assigned to the second stimulus and several intermediate weightings between these two extremes. See "Simulation" section.

the animal would choose to repeat the stimulus more frequently when the stimulus is difficult. In addition, metacognition predicts that the animal would perform more accurately when choosing to take the test immediately compared to trials in which it is forced to take the test immediately.

Application of the response-strength model (Fig. 1) to the procedure (Fig. 2) proceeds as follows for difficult stimulus conditions (i.e., near the middle of the x-axis in Fig. 1); in this application, the flat threshold in Fig. 1 is referred to as the repeat response. Typically, response strength for short and long is low for difficult problems relative to the repeat response. Hence, repeating the stimulus should occur frequently for difficult problems. However, on some rare occasions, response strength for short or long will exceed the repeat-response threshold. On those occasions, the subject is expected to choose to take the test immediately. Because these are difficult problems, performance would be *relatively poor* on these trials in which the subject chooses to take the test immediately (i.e., there is an oversampling of poor performance on choice-take trials). Importantly, the low-performing trials contribute to the choice-take trials and do *not* contribute to the accuracy tests that terminate choice-repeat trials (i.e., there is an under-sampling of poor performance on choice-repeat trials); because a repeat-response threshold is most likely to exceed the response strength for short or long durations for the most difficult problems, these repeat responses under-sample poor performance on choice-repeat trials more on difficult problems than on easy problems. Therefore, accuracy on choice-repeat trials is increased (because they do not include some of the low-performing trials as outlined above). By contrast, forced-repeat trials include both the low-performing trials and those with higher performance. The combination of low and high levels produces an average level of performance that is low relative to choice-repeat performance; note that this is opposite to the prediction developed in the paragraph above based on metacognition.

Methods

Subjects—Eight male Sprague–Dawley rats (*Rattus norvegicus*; Harlan, Madison, WI; 85 days old) were individually housed in a colony room with a reversed 12–12 light–dark schedule (light offset at 07:00, onset at 19:00). Testing began when the rats were approximately 131 days old and weighed an average of 277 g. During pretraining and testing sessions rats received 45-mg pellets (F0165, Bio-Serv, Frenchtown, NJ) and later received a supplemental ration of 5001-Rodent-Diet (Lab Diet, Brentwood, MO) for a total daily ration of 15–20 g. Water was available continuously. All procedures were approved by the University of Georgia institutional animal care and use committee and followed the guidelines set forth by the National Research Council *Guide for the Care and Use of Laboratory Animals*.

Pilot study: Rats were trained to discriminate short and long stimulus durations (see "duration discrimination" below) and to use nose-pokes. As described in the Introduction, a central prediction applies to difficult problems in our design. Terminal performance from the initial duration-discrimination training was used as a baseline level of performance to identify the difficult stimulus for each rat. Next, rats received forced-repeat, forced-test, and choice trials, as outlined in Fig. 2 and described in greater detail below. Note that the baseline and repeat the stimulus data were collected in different stages of the pilot. Thus, one limitation of this approach in the pilot study was that the identification of a difficult stimulus was fixed and was based on increasingly old baseline data as the data from the repeat-the-stimulus task were obtained. In addition, an animal may adjust its duration criterion slightly from day-to-day. Therefore, the approach developed in the pilot was refined to train the rats on the duration discrimination in the first half of the daily session, followed by the repeat-the-stimulus task for the remainder of the session (as described

below). This approach permitted a daily estimate of the difficult stimulus to ensure that the use of repeat-the-test and take-the-test responses could be isolated to currently difficult stimuli. The total number of sessions for the pilot study varied (due to individual differences in learning) according to subject as follows: KK1 and KK7 completed 11 sessions; KK2, KK3, KK4, KK6, and KK8 completed 21 sessions; and KK5 completed 14 sessions. The maximum number of trials was approximately 33 per session.

Apparatus—Eight identical operant chambers ($30 \times 28 \times 23$ cm, width × height × depth; Med Associates ENV-007, Georgia, VT), each located within a ventilated sound-attenuation cubicle (ENV-016 M, $66 \times 56 \times 36$ cm, W × H × D), were used for the experiment. Each operant chamber contained a recessed food trough (ENV-200R2 M, 5×5 cm) equipped with a photobeam (used to detect head entries; ENV254, 1 cm in from food trough, 1.5 cm from bottom of food trough) that was centered horizontally (63 cm above the floor) between two retractable levers (ENV-112CMX) on one wall of the chamber. A 45-mg pellet dispenser (ENV-203-45IRX) was located on the outside wall of the chamber and was attached to the food trough. A photobeam located on the pellet dispenser detected successful pellet delivery. A pellet dispenser would make up to four additional attempts to dispense pellets if a failure was detected. A water bottle with an attached sipper tube was placed on the outside wall opposite the food trough. The sipper tube was inserted into the chamber via a 1×1.5 cm opening in the wall. A nose-poke aperture (2.5 cm diameter) was located to either the left or the right of the sipper tube and contained a photobeam that detected individual entries. A retractable automated guillotine door (ENV-210M) was used to give/restrict access to each nose-poke opening. The floor of the chamber consisted of 19 stainless steel rods (4 mm diameter, 15.5 mm spacing), and a stainless steel waste tray was located below the chamber floor. Other equipment included a clicker (ENV-135M), lights (ENV-215M and ENV-227M), a speaker (ENV-225SM), a photobeam lickometer (ENV-251L), and four equally spaced photobeams that were 4 cm above the floor. A computer with a Celeron processor (850 MHz) running Med-PC (version 4.0) was located in a nearby room and controlled experimental events and recorded the time at which each event occurred (10 ms accuracy).

Procedure

Pretraining: Rats were given feeder training that consisted of one food pellet being delivered per minute, accompanied by a click before pellet delivery, for one 30-min session. Next, rats underwent 60-min daily sessions of lever and nose-poke training for 3 and 4 days, respectively, in which each pellet was contingent a single response.

Duration-discrimination training: The stimuli that were used for the duration-discrimination training were eight logarithmically spaced white-noise durations: 2.00, 2.44, 2.97, 3.62, 4.42, 5.38, 6.56, and 8.00 s. Stimuli were chosen by independent random selection before the start of each trial. Rats were trained to discriminate short and long noise durations. Short durations ranged from 2.00 to 3.62 s, and long durations ranged from 4.42 to 8.00 s. Duration discrimination became more difficult as a stimulus approached 4.00 s (i.e., the easiest durations to discriminate were 2.00, 2.44, 2.97, 5.38, 6.56, and 8.00 s while the most difficult durations to discriminate were 3.62 and 4.42 s). The inter-trial interval (ITI) was 8–10 min for each 9-h daily session. A trial began with the presentation of a 70-dB white-noise stimulus duration that rats had to classify as either short or long. Rats indicated their choice by pressing one lever for short and one lever for long (lever assignment was counterbalanced across subjects prior to the beginning of the experiment). Rats received a large reward of six pellets for correctly classifying the duration as short or long and received no reward for incorrectly discriminating a stimulus. Duration-discrimination training continued until each subject achieved an average accuracy score of at least 75% across all

eight stimulus durations for at least three consecutive sessions. Rats were initially trained with a non-correction procedure (as described above). To facilitate accuracy, a correction procedure was introduced next; a trial with a randomly selected duration was repeated until a correct duration-classification response occurred. Next, the rats were returned to the non-correction procedure, and non-correction procedures were used throughout the remainder of the experiment.

Metacognition testing

Repeat the stimulus procedure: Metacognition was assessed by using the repeat-the-stimulus procedure that was developed as a refinement of the procedure used in the pilot study. The repeat-the-stimulus procedure established a daily estimate of the most difficult stimulus for each subject. Each daily session was comprised of two parts: (1) In the first part of each daily session, rats received training on the duration discriminations, (2) In the second part of the session, rats proceeded through the repeat-the-stimulus task (see Fig. 2). The transition between the two parts occurred approximately halfway through the procedure and varied randomly across subjects and days. The maximum number of trials was approximately 36 per session.

Each repeat-the-stimulus trial consisted of three phases. During Phase 1, a white-noise stimulus duration was presented and would later be classified in Phase 3. During Phase 2, the rat was either forced to take a duration test (1/3 of trials), forced to repeat the stimulus (1/3 of trials), or had the choice (1/3 of trials) to take a duration test or repeat the stimulus. In Phase 3, rats pressed one lever to identify the stimulus as short or pressed the other lever to identify the stimulus as long. If the rat correctly classified the stimulus duration, it received a large reward of six pellets. However, if it classified the stimulus duration incorrectly, it received no reward. Rats received an additional small, one pellet reward immediately upon choosing or being forced to choose to repeat the stimulus. In research by Roberts et al. (2009), pigeons had the option to request access to a sample stimulus (without immediate reward) in a matching to sample experiment, but the pigeons virtually never selected the study response; hence, to offset the more attractive delay to reinforcement available when the rat chose to take an immediate test, we chose to provide immediate reward, thereby encouraging the rats to sometimes sample the repeat response. Trial type was randomly selected prior to the start of each trial. The stimulus duration was chosen by independent random selection before the start of each trial. Assignment of nose-pokes to take and repeat conditions and assignment of levers to short and long correct responses was counterbalanced across subjects prior to the start of the experiment. Rats were tested using three different trial types (outlined in the "Conditions" section below) until their average accuracy level was greater than or equal to 75% for the four easiest short and long stimulus durations (easy short = 2.00 and 2.44; easy long = 6.56 and 8.00).

Conditions

Forced-test trials: On forced-test trials, rats were forced to take a stimulus duration test. Forced-test trials began with the presentation of a white-noise stimulus duration (i.e., Phase 1, see left side of Fig. 2). In Phase 2, the guillotine door covering the "take the test" nose-poke retracted and allowed access only to the "take the test" nose-poke (e.g., the left nose-poke in Fig. 2). After the guillotine door retracted, rats were required to break the photobeam in the "take the test" nose-poke aperture (e.g., the left nose-poke in Fig. 2) in order to move to Phase 3. As soon as the rat broke the photobeam in the nose-poke, the guillotine door closed. In Phase 3 levers inserted into the chamber, and rats were required to press one lever. If the rat classified the stimulus duration correctly, it received a reward of six pellets, and if it was incorrect it received no pellets.

Forced-repeat trials: On forced-repeat trials, rats were forced to repeat a stimulus (i.e., the same stimulus duration was presented again), which was later followed by a forced-test. Forced-repeat trials began with the presentation of a white-noise stimulus duration (i.e., Phase 1, see right side of Fig. 2). In Phase 2, rats were forced to hear a re-presentation of the same stimulus duration presented during Phase 1. Phase 2 began with the retraction of the guillotine door covering the "repeat the stimulus" nose-poke (e.g., the right nose-poke in Fig. 2). Only the "repeat the stimulus" nose-poke was accessible and the guillotine door on the other nose-poke remained closed. Rats were required to break a photobeam in the "repeat the stimulus" nose-poke in order to hear a re-presentation of the same stimulus duration presented in Phase 1 and receive a 1 pellet reward. After rats heard the stimulus duration for a second time, the remainder of the trial proceeded in the same manner as a forced-take trial.

Choice trials: In choice trials, rats had the opportunity to choose to take the duration test (i.e., choice-take) or to hear a re-presentation of the stimulus (i.e., choice-repeat). Choice trials began with the presentation of a white-noise stimulus duration (i.e., Phase 1, center of Fig. 2). Afterward, in Phase 2, both guillotine doors covering the "take the test" and the "repeat the stimulus" nose-pokes retracted, which allowed the rats to access both nose-pokes. If rats chose to take a duration test, the remainder of the trial proceeded as in a forced-take trial. On the other hand, if rats chose to repeat the stimulus the remainder of the trial proceeded as in a forced-repeat trial.

Data analysis: The total number of sessions for the meta-cognition testing procedure differed for individual rats. Fifteen sessions were omitted for all subjects due to an equipment problem. Subject KK1 exhibited a response bias on the duration-discrimination task after the equipment problem and did not progress to subsequent testing procedures because its performance did not exceed 75% correct for the four easiest short and long stimulus durations (easy short = 2.00 and 2.44; easy long = 6.56 and 8.00). As a result, subject KK1 did not receive the metacognition testing procedure.

Data from each daily session were divided into two parts. Data from the first part of each daily session were used to estimate accuracy for the difficult stimuli and for the easy stimuli. The first part of each daily session identified the stimulus that each subject found most difficult by determining whether the accuracy for each stimulus duration was less than 75% correct. To estimate accuracy for the easy stimuli, the following stimulus durations were used: 2.00, 2.44, 6.56, and 8.00 s. To estimate accuracy for the difficult stimuli, the following durations were used: 3.62 and 4.42 s. These two stimulus durations were selected based upon what subjects found to be the most difficult stimulus on each day. Data from the second part of each daily session were then examined for easy and difficult conditions using the stimuli identified for each subject. Proportion correct was calculated by dividing the total number of correct trials by the total number of trials for each stimulus duration separately for each trial type (i.e., choice-repeat or forced-repeat). The analyses were performed on the 70 terminal sessions for all subjects (70 sessions was the minimum number of sessions completed by all subjects).

Results

If rats have knowledge about their own cognitive states, then rats should be more accurate when forced to repeat a difficult stimulus duration test than when they have the option to do so on difficult trials. Figure 3 shows accuracy as a function of difficulty for both choice-repeat and forced-repeat trials. As expected, accuracy is lower on difficult trials compared to easy trials. Importantly, choice-repeat accuracy is lower than forced-repeat accuracy on difficult trials. A two-factor analysis of variance (ANOVA) was performed on accuracy for

trial type (i.e., choice-repeat vs. forced-repeat trials) and stimulus difficulty (i.e., easy vs. difficult). The ANOVA revealed a significant main effect of stimulus difficulty, $F(1, 6) = 133.05$, $P = 0.00003$, and a significant main effect of trial type, $F(1, 6) = 13.68$, $P = 0.01$. Additionally, the ANOVA revealed a significant interaction between stimulus difficulty and trial type, $F(1, 6) = 6.98$, $P = 0.038$. A planned comparison of accuracy on difficult stimuli for forced-repeat trials and choice-repeat trials was analyzed using a paired-samples t test ($N = 7$, $M = -0.051$, SEM = 0.012; the mean and standard error of the mean associated with t tests throughout this article report descriptive statistics for difference scores). As predicted, rats were more accurate on trials in which they were forced to repeat a difficult stimulus duration test than on trials in which they chose to repeat a difficult test, $t(6) = -4.39$, $P = 0.005$. The accuracy difference in Fig. 3 was not produced by a longer retention interval in choice-repeat relative to forced-repeat trials. The latency to respond at the repeat-the-stimulus nose-poke in forced (1.19 ± 0.07 s; mean ± SEM) and choice (1.16 ± 0.08 s) conditions was similar and not significantly different, $t(6) = 0.5$, $P = 0.65$.

On choice trials, rats would be expected to repeat the stimulus more often on difficult stimulus durations than on easy stimulus durations according to both metacognition and non-metacognition proposals. The observed rate of choosing to repeat the stimulus was 0.638 and 0.671 for easy and difficult conditions, respectively. A paired-samples t test ($N = 7$, $M = 0.033$, SEM = 0.022) did not reveal a significant difference in the frequency of choosing to repeat the stimulus more often on difficult stimulus durations than on easy stimulus durations, $t(6) = 1.49$, $P = 0.187$. The data are consistent with the predicted direction for both metacognition and non-metacognition proposals (i.e., rats had a very small tendency to repeat the stimulus more often on difficult stimulus durations) although the difference was not statistically significant.

Although metacognition and the response-strength model do not make differential predictions in other conditions, we have provided some additional information to describe other aspects of performance on the task (see Table 1 for descriptive statistics); the other aspects of performance in Table 1 come from the parts of the procedure in which the subjects took the test immediately (i.e., they did not repeat the stimulus). Accuracy on choice-take trials is expected to be higher than accuracy on forced-take trials according to both metacognition and non-metacognition proposals. A paired-samples t test ($N = 7$, $M = -0.11$, SEM = 0.084) was used to compare accuracy on forced-take trials with accuracy on choice-take trials. Accuracy between forced-take trials and choice-take trials was not significantly different for difficult stimuli, $t(6) = -1.36$, $P = 0.22$. In order to determine whether hearing the stimulus duration for a second time improved performance on the stimulus discrimination, the average accuracy for forced-take trials and choice-take trials (take condition) was compared to accuracy on forced-repeat trials for both easy and difficult stimuli. A paired-samples t test ($N = 7$, $M = -0.09$, SEM = 0.052) revealed a trend toward a significant difference between accuracy for the take condition and forced-repeat trials for difficult stimuli, $t(6) = -1.74$, $P = 0.06$ (one-tailed). For easy stimuli, a paired-samples t test ($N = 7$, $M = -0.08$, SEM = 0.036) revealed a significant difference between accuracy for the take condition and forced-repeat trials, $t(6) = -2.29$, $P = 0.03$ (one-tailed). These results provide some limited support that two stimulus presentations were more beneficial than one stimulus presentation. Use of one-tailed tests are appropriate in these two cases because the prediction is that two stimuli are better than one stimulus presentation, which is a directional prediction.

In some studies of metacognition, some subjects produce a pattern of data consistent with metacognition, while other subjects do not (e.g., Beran et al. 2009). All of the above analyses were averaged across all subjects. Next, we examined the individual data to see if any single subject produced the complete pattern of data predicted by metacognition. One

rat (KK7) is a close match to the meta-cognition predictions. Accuracy was higher in forced-repeat (0.679) than in choice-repeat (0.574) trials (by 0.105). Accuracy was higher in choice-take (1.00) than in forced-take (0.632) trials (by 0.368). The relative frequency of choice-repeat responses was higher on difficult (0.978) than on easy (0.927) stimuli (by 0.051). A yield of one out of eight rats is rather low, which may happen by the post hoc nature of examining individual data for a particular pattern.

Simulation

The objective of the simulation was to determine if the response-strength model could produce the accuracy difference predicted by metacognition and observed in Fig. 3. Simulations were conducted to determine whether the response-strength model could fit the data. Although we focus on the Smith et al. (2008) model outlined in Fig. 1, it is important to note that other non-metacognition hypotheses have been offered. Smith et al. (2008) described two non-metacognition proposals and Staddon et al. (Jozefowiez et al. 2009a; Jozefowiez et al. 2009b) described additional alternatives. Each proposal has a similar function to model the decision process. Thus, we examine in detail one of Smith et al.'s proposals. Other proposals are qualitative rather than quantitative (Hampton 2009). We believe that our conclusions could be similarly derived using other versions of non-metacognition proposals.

Method

The simulation began with an exhaustive search of the parameter space in order to identify the least-squares best fitting parameters (i.e., the parameters that minimized the sum of squared deviations between the empirical data and the simulated values). A minimum, maximum, and step-size was used for each parameter. The parameters that minimized the difference between the data and the simulation were identified.

The simulation closely followed the "Play it Again" procedure and therefore, the process of identifying parameters for the simulation is described in procedural terms. The range of stimulus durations (2–8 s) was expressed as values within the range of 1–71 [following Smith et al. (2008) simulations]. An objective physical stimulus is perceived with variability. This feature was modeled by sampling from a normal distribution with a mean that corresponds to the objective physical stimulus (stimulus mean) and a parameter for the standard deviation of the distribution (stimulus SD). Thus, a subjective duration was determined on each simulated trial by a random number, a mean, and a standard deviation. The response strength to judge the subjective duration as short or long was determined by an exponential curve (see Fig. 1), a sensitivity parameter for the exponent (sens), and the subjective duration described above (errs; i.e., the distance between the subjective and anchor durations). The exponential curve was calculated using Smith et al. (2008) equation $e^{-\text{sens}\times\text{errs}}$ (p. 691). The decision to repeat the stimulus or take an immediate test was modeled by a flat response threshold (i.e., a constant level of attractiveness, independent of the magnitude of the objective physical stimulus). The value for response threshold was modeled by sampling from a normal distribution with a mean (threshold mean) and standard deviation (threshold SD) as parameters.

To simulate the repeat-the-stimulus condition, each stimulus presentation was modeled with an independent random sample using a set of parameters with random variation. The subjective duration after two stimulus presentations was modeled by a weighted average of the two independent stimulus presentations (Weighted Average). A winner-take-all response rule was applied to response strengths for the short, long, and repeat-the-stimulus responses. The duration-classification response was based on a winner-take-all response rule for short and long. Impossible values (e.g., durations below zero) were discarded and resampled.

Accuracy is based on averaging (i.e., the relative frequency of) outcomes for incorrect (represented as 0) and correct (represented as 1) outcomes.

Each simulation consisted of 10,000 trials. The parameter set for an individual simulation was selected from the minimum, maximum, and step size values shown in Table 2. An exhaustive search of the 282, 240 sets of parameters was conducted for easy (corresponding to durations 2.00 and 2.44 s; stimulus categories 1 and 11 in the simulation) and difficult (corresponding to 3.62 s; stimulus category 31 in the simulation) conditions. Because difficulty in the duration task is indexed by the distance between stimulus duration and the point of subjective equality of 4.0 s, we simulated conditions below 4.0 s. Duplication of equivalent levels of difficulty above 4.0 s would have increased the number of simulations to over 2.5 million simulations, which was deemed prohibitive with respect to the computer resources. The simulation outcomes for the easy conditions were averaged (as was done for the data) at each parameter set.

In each simulation, six values were estimated: the proportion of trials in which a choice to repeat the stimulus occurred in easy and difficult conditions; proportion correct on forced-repeat trials in easy and difficult conditions; and proportion correct on choice-repeat trials in easy and difficult conditions. The sum of squared differences between observed (data) and expected (simulation) values was calculated for each set of parameters using the six proportions listed above. The set of parameters that minimized the sum of squared deviations is presented in Table 3. The set of parameters identified by the least-squares method described above was examined further as follows. Seven new simulations (each consisting of 10,000 trials) were performed to estimate variability across seven simulated subjects, which corresponds to the sample size in the data.

Results

The goal of the simulation was to determine whether the response-strength model could fit the observed accuracy difference in the experimental data. The exhaustive search of the parameter space identified a set of parameters that minimized the error between the empirical data (see Fig. 3) and simulated values. The results of the simulation are shown in Fig. 4. Unlike the experimental data, the simulated data do not show a difference in accuracy between choice-repeat trials and forced-repeat trials for difficult stimuli, $t(6) = 0.85$, $P = 0.4$ ($N = 7$, $M = 0.0016$, SEM = 0.0018). According to the results of the simulation, we would expect no significant difference in accuracy on choice-repeat trials and forced-repeat trials for difficult stimulus durations, which is contrary to the data shown in Fig. 3.

Importantly, the empirical data produced a statistically significant difference between forced-repeat and choice-repeat accuracy in the difficult conditions (see Fig. 3). Consequently, we wanted to determine whether the magnitude of this difference was statistically different from the expected difference according to the response-strength simulation. An independent samples t test ($M = 0.0493$, SEM = 0.0117) was conducted on the empirical data and the simulated data. The magnitude of the accuracy difference in forced-repeat and choice-repeat conditions was larger in the empirical data than in the simulation, $t(12) = -4.21$, $P = 0.001$. Hence, the empirical data showed an accuracy difference that could not be explained by the response-strength model.

Discussion

The objective of the experiment and simulation was to develop a method for testing metacognition in animals that could not be explained by a response-strength model. Our data showed the predicted difference in accuracy on choice-repeat and forced-repeat trials (Fig. 3). In contrast, the response-strength model does not predict an accuracy difference on

choice-repeat and forced-repeat trials. One potential explanation for the observed accuracy difference is that making a choice took more time than the forced-choice condition. Degraded accuracy due to a relatively long retention interval is unlikely to explain our data because the subject-imposed retention intervals did not differ between choice-repeat and forced-repeat trials. Nevertheless, it is not known if a different substantive choice would have also decreased accuracy. However, it is not clear why a substantive choice would selectively degrade accuracy on difficult, but not easy, stimuli.

We performed simulations of the "Play it Again" procedure to determine whether the response-strength model could fit the accuracy difference between forced-repeat and choice-repeat trials. Specifically, we conducted an exhaustive search of parameters and found that the simulated data predicted equal performance on difficult stimulus discriminations for both forced-repeat and choice-repeat trials. In contrast, our experimental data showed a significant accuracy difference between choice-repeat and forced-repeat trials as predicted by metacognition.

Although the data show the accuracy difference predicted by metacognition for forced-repeat and choice-repeat trials, the data do not show other features that would be expected by metacognition. Both metacognition and the response-strength model predict that rats should choose to repeat difficult stimulus discriminations more often than easy stimulus discriminations. Metacognition predicts the repeat response because the subject is assumed to know that he does not know the answer to the difficult problem. However, the same prediction is made without hypothesizing metacognition. Repeated presentation of a difficult stimulus would increase accuracy, and thereby increase reward yield. Hence, a stimulus–response rule would allow a subject to learn to repeat the stimulus to obtain more reward. Our results suggest that rats had a small numerical tendency for choosing to repeat the stimulus more often for difficult stimulus durations. Although this finding is consistent with the direction of our prediction for both metacognition and the response-strength model, it was not a statistically significant difference. Reinforcement of the repeat nose-poke response (see Fig. 2) may have masked the expected difference in take versus choice nose-poke responses. To encourage the rats to sample the repeat nose-poke response, a pellet reward was delivered contingent on the repeat nose-poke response; this reinforcement was needed to offset the more attractive delay to reinforcement available when the rat chose to take an immediate test. Therefore, we suspect that the design of the procedure allowed for greater sensitivity for detecting differences in duration-discrimination accuracy than for choice of differentially reinforced nose-poke responses.

It is surprising that the rats showed the accuracy difference predicted by metacognition while not showing a reliable preference for the choice-repeat option on difficult trials. One potential explanation is that choice-repeat responses are over-represented because the animals were influenced by the reward associated with that response. Over-representation of choice-repeat response may decrease the magnitude of the observed accuracy difference. Hence, a larger accuracy difference might be observed if the influence of direct reward for the choice-repeat response could be curtailed. The contingency between the repeat-the-stimulus response and the subsequent pellet may have also differentially recruited attention toward the subsequently repeated duration. An independent reason to believe that the immediate reward impacted choice responses comes from a comparison with similar experiments by Roberts et al. (2009) with pigeons. In that research, pigeons had the option to request access to a sample stimulus in a matching to sample experiment, but the pigeons virtually never selected the study response; by contrast, our rats selected the repeat option on about 65% of the trials. We used a single pellet as a minimum reward. One method that future research can use to balance the use of a choice-repeat response is to explore the range

between our one pellet and Roberts et al. zero pellet conditions by presenting the pellet probabilistically and titrating the probability for individual subjects.

In addition, the rats did not show an accuracy difference in forced-take versus choice-take trials, contrary to the predictions of both metacognition and response-strength models. Although the difference in accuracy with difficult stimuli was not significantly different, accuracy was numerically lower on choice-take tests compared to forced-take condition. One factor that may lower accuracy in choice-take trials is the presence of a decision (i.e., a choice), which was absent in the forced-take condition. Overall, the evidence for metacognition in rats is inconsistent, with one feature uniquely predicted by metacognition confirmed but with two other features predicted by both models not con-firmed. Clearly, putative evidence for metacognition in rats is critically undermined when part of the data are not consistent with metacognition. Consequently, the current data do not support the hypothesis that rats exhibit metacognition.

We expect that some refinements to our procedure can be introduced in future research, but we emphasize that it is important that these be carefully tested against simulations of a response-strength model. For example, it would be interesting to force the animal to take the test by surprise after it chooses to repeat and to force the animal to take the test immediately without offering a choice to repeat or take the test. It would also be valuable to decrease the impact of reward on the repeat response by using a relatively low probabilistic reward.

The development of animal models of metacognition holds enormous potential for understanding the neuroanatomical, neurophysiological, neurochemical, and genetic mechanisms of metacognition. Moreover, an animal model may ultimately contribute insight into cognitive impairments (e.g., understanding the failure of the distinctiveness heuristic in Alzheimer's disease, understanding how changes in brain morphology affect rumination in patients with depression, and alleviating the effects of nicotine withdrawal on memory and metacognition; Budson et al. 2005; Kelemen and Fulton 2008; Roelofs et al. 2007). Additionally, investigating metacognition in animals may provide insight into the evolution of the mind (Emery and Clayton 2001; Hampton 2001; Kornell 2009; Son and Kornell 2005; Smith 2005, 2009).

Conclusions

We believe that the "Play it Again" method and the simulations represents progress in developing valid methods for testing metacognition in animals. The simulations suggest that the response-strength model does not predict the accuracy difference between choice-repeat trials and forced-repeat trials in our data. This prediction is contrary to the prediction made by metacognition and to the accuracy difference observed in our data. However, other aspects of performance were not consistent with metacognition. Consequently, our findings provide inconsistent support for metacognition in rats. The current results call into question previous findings suggesting that rats have metacognitive abilities (Foote and Crystal 2007). However, we believe that the play it again method combined with simulations may be valuable for testing metacognition in other species.

Our findings demonstrate that valid methods for testing metacognition in animals can be developed. Importantly, performing simulations required us to develop our new method and precisely specify the predictions made by non-metacognitive explanations. Although our method was designed to test metacognition in rats, an important next step is to use similar methods (i.e., a combination of novel procedures and simulations) with other animals. Using similar methods with other animals would provide converging lines of evidence for metacognition in other animals. We believe the "Play it Again" method and simulations

have the potential to resolve controversies about the existence of metacognition in nonhuman animals.

Acknowledgments

References

Basile BM, Hampton RR, Suomi SJ, Murray EA. An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*). Anim Cogn. 2009; 12:1169–1180.

Beran MJ, Smith JD, Coutinho MVC, Couchman JJ, Boomer J. The psychological organization of "uncertainty" responses and "middle" responses: a dissociation in capuchin monkeys (*Cebus apella*). J Exp Psychol Anim Behav Proc. 2009; 35:371–381.

Budson AE, Dodson CS, Daffner KR, Schacter DL. Metacognition and false recognition in Alzheimer's disease: further exploration of the distinctiveness heuristic. Neuropsychology. 2005; 19:253–258. [PubMed: 15769209]

Call J. Do apes know that they could be wrong? Anim Cogn. 2010; 13:689–700. [PubMed: 20306102]

Call J, Carpenter M. Do apes and children know what they have seen? Anim Cogn. 2001; 4:207–220.

Cardinal RN, Pennicott DR, Sugathapala CL, Robbins TW, Everitt BJ. Impulsive choice induced in rats by lesions of the nucleus accumbens core. Science. 2001; 292:2499–2501. [PubMed: 11375482]

Church RM, Deluty MZ. Bisection of temporal intervals. J Exp Psychol Anim Behav Proc. 1977; 3:216–228.

Crystal JD, Foote AL. Metacognition in animals. Comp Cogn Behav Rev. 2009; 4:1–16.

Emery NJ, Clayton NS. Effects of experience and social context on prospective caching strategies by scrub jays. Nature. 2001; 414:443–446. [PubMed: 11719804]

Foote AL, Crystal JD. Metacognition in the rat. Curr Biol. 2007; 17:551–555. [PubMed: 17346969]

Gibbon J. On the form and location of the psychometric bisection function for time. J Math Psychol. 1981; 24:58–87.

Hampton RR. Rhesus monkeys know when they remember. Proc Natl Acad Sci USA. 2001; 98:5359–5362. [PubMed: 11274360]

Hampton RR. Multiple demonstrations of metacognition in non-humans: converging evidence or multiple mechanisms? Comp Cogn Behav Rev. 2009; 4:17–28. [PubMed: 20046911]

Hampton RR, Zivin A, Murray EA. Rhesus monkeys *(Macaca mulatta)* discriminate between knowing and not knowing and collect information as needed before acting. Anim Cogn. 2004; 7:239–246. [PubMed: 15105996]

Inman A, Shettleworth SJ. Detecting metamemory in nonverbal subjects: a test with pigeons. J Exp Psychol Anim Behav Proc. 1999; 25:389–395.

Jozefowiez J, Cerutti DT, Staddon JER. The behavioral economics of choice and interval timing. Psychol Rev. 2009a; 116:519–539. [PubMed: 19618985]

Jozefowiez J, Staddon JER, Cerutti DT. Metacognition in animals: how do we know that they know? Comp Cogn Behav Rev. 2009b; 4:29–39.

Kelemen WL, Fulton EK. Cigarette abstinence impairs memory and metacognition despite administration of 2 mg nicotine gum. Exp Clin Psychopharmacol. 2008; 16:521–531. [PubMed: 19086773]

Kornell N. Metacognition in humans and animals. Curr Dir Psychol Sci. 2009; 18:11–15.

Kornell N, Son L, Terrace H. Transfer of metacognitive skills and hint seeking in monkeys. Psych Sci. 2007; 18:64–71.

Mazur JE. Choice between small certain and large uncertain reinforcers. Anim Learn Behav. 1988; 16:199–205.

Mazur J. Rats' choices between one and two delayed reinforcers. Learn Behav. 2007; 35:169–176. [PubMed: 17918422]

McMahon S, Macpherson K, Roberts WA. Dogs choose a human informant: metacognition in canines. Behav Process. 2010; 85:293–298.

Metcalfe, J.; Kober, H. Self-reflective consciousness and the projectable self. In: Terrace, HS.; Metcalfe, J., editors. The missing link in cognition: origins of self-reflective consciousness. Oxford University Press; New York: 2005. p. 57-83.

Richards J, Mitchell S, de Wit H, Seiden L. Determination of discount functions in rats with an adjusting-amount procedure. J Exp Anal Behav. 1997; 67:353–366. [PubMed: 9163939]

Roberts WA, Feeney MC, McMillan N, MacPherson K, Musolino E, Petter M. Dopigeons (*Columba livia*) study for a test? J Exp Psychol Anim Behav Proc. 2009; 35:129–142.

Roelofs J, Papageorgiou C, Gerber RD, Huibers M, Peeters F, Arntz A. On the links between self-discrepancies, rumination, metacognitions, and symptoms of depression in undergraduates. Behav Res Ther. 2007; 45:1295–1305. [PubMed: 17156745]

Shepard RN. Application of a trace model to the retention of information in a recognition task. Psychometrika. 1961; 26:185–203.

Shepard RN. Toward a universal law of generalization for psychological science. Science. 1987; 237:1317–1323. [PubMed: 3629243]

Smith, JD. Studies of uncertainty monitoring and metacognition in animals and humans. In: Terrace, HS.; Metcalfe, J., editors. The missing link in cognition: origins of self-reflective consciousness. Oxford University Press; New York: 2005. p. 242-271.

Smith JD. The study of animal metacognition. Trends Cogn Sci. 2009; 13:389–396. [PubMed: 19726218]

Smith JD, Beran MJ, Redford JS, Washburn DA. Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. J Exp Psychol Gen. 2006; 135:282–297. [PubMed: 16719654]

Smith JD, Beran MJ, Couchman JJ, Coutinho MVC. The comparative study of metacognition: sharper paradigms, safer inferences. Psychon Bull Rev. 2008; 15:679–691. [PubMed: 18792496]

Smith JD, Redford JS, Beran MJ, Washburn DA. Rhesus monkeys (*Macaca mulatta*) adaptively monitor uncertainty while multi-tasking. Anim Cogn. 2010; 13:93–101. [PubMed: 19526256]

Sole LM, Shettleworth SJ, Bennett PJ. Uncertainty in pigeons. Psychon Bull Rev. 2003; 10:738–745. [PubMed: 14620372]

Son, LK.; Kornell, N. Metaconfidence judgments in rhesus macaques: explicit versus implicit mechanisms. In: Terrace, HS.; Metcalfe, J., editors. The missing link in cognition: origins of self-reflective consciousness. Oxford University Press; New York: 2005. p. 296-320.

Sutton JE, Shettleworth SJ. Memory without awareness: pigeons do not show metamemory in delayed matching-to-sample. J Exp Psychol Anim Behav Proc. 2008; 34:266–282.

Washburn DA, Gulledge JP, Beran MJ, Smith JD. With his memory magnetically erased, a monkey knows he is uncertain. Biol Lett. 2010; 6:163–166. [PubMed: 19864274]

White KG. Psychophysics of remembering: the discrimination hypothesis. Curr Dir Psychol Sci. 2002; 11:141–145.
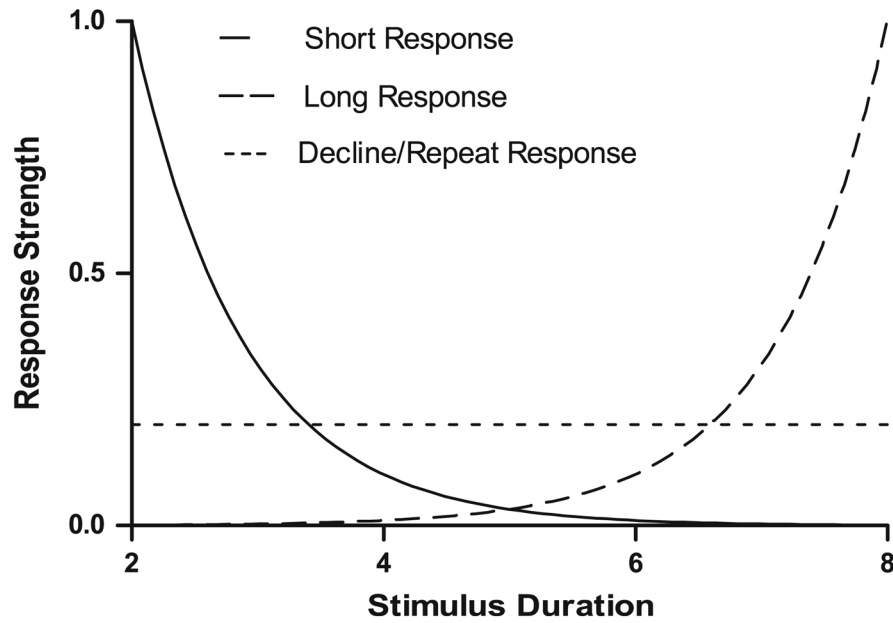
Fig. 1.
Smith et al. (2008) associative model of stimulus generalization/response strength. Smith et al. developed a model of the putative metacognitive data pattern by using exponential curves, response strength, and a reward threshold (i.e., a decline response) as a function of the subjective impression of the stimulus. One exponential curve represents the response strength of a short stimulus duration while the other exponential curve represents the response strength of a long stimulus duration. Both exponential curves cross one another as response strengths decrease. Additionally, there is a relatively low and flat threshold for the decline response. Because the reward threshold is above the point where the two exponential curves cross, the decline response is selected for trials that would otherwise generate low accuracy. Importantly, this model predicts the emergence of accuracy divergence because poor-performing trials are selectively removed when tests are declined on difficult trials in the experimental design by Foote and Crystal (2007). When the figure is applied to the current experiment (see Fig. 2), the flat threshold is referred to as the repeat response
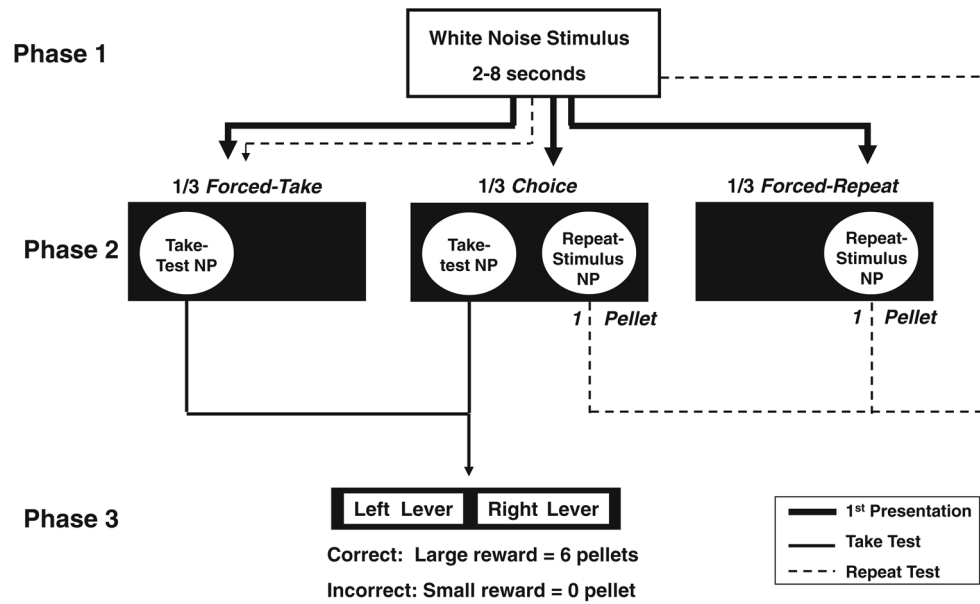
Phase 1

White Noise Stimulus
2-8 seconds

1/3 *Forced-Take*   1/3 *Choice*   1/3 *Forced-Repeat*

Phase 2

Take-Test NP

Take-test NP   Repeat-Stimulus NP   Repeat-Stimulus NP

1 *Pellet*   1 *Pellet*

Phase 3

Left Lever   Right Lever

Correct: Large reward = 6 pellets
Incorrect: Small reward = 0 pellet

1st Presentation
Take Test
Repeat Test

Fig. 2.
Schematic representation of the "Play it Again" method. Nose-poke is abbreviated as NP.
Each trial consisted of three phases. In phase 1, rats were presented with a white-noise
stimulus duration. In phase 2, rats were forced to either take a duration test, repeat the
stimulus from phase 1, or were given the choice to take a test or repeat the stimulus. In
phase 3, rats used levers to classify the stimulus from phase 1 as short and long. The
stimulus duration and trial type were randomly selected before the beginning of each trial.
Bold lines indicate one of three ways in which a trial could proceed. *Thin solid lines* indicate
the sequence of events for a forced-take (*left side of diagram*) or a choice-take trial (*center
of diagram*). *Dashed lines* indicate the two ways a trial could be repeated, a choice-repeat
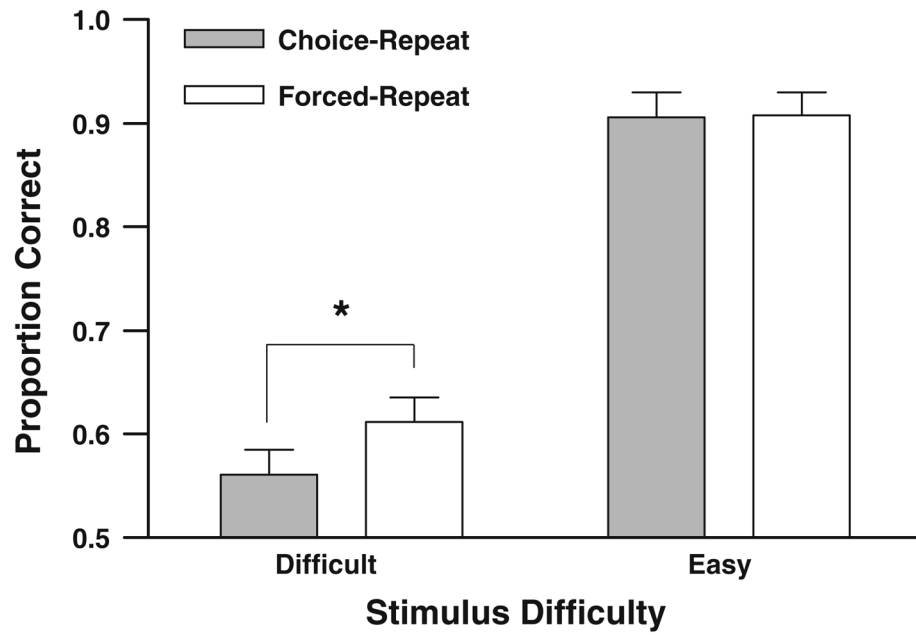trial (*center of diagram*) or a forced-repeat trial (*right side of diagram*)

Fig. 3.
Accuracy of rats on each trial type as a function of stimulus difficulty. Difficult stimuli included stimulus durations 3.63 and 4.42 s. Easy stimuli included stimulus durations 2.00, 2.44, 6.56, and 8.00 s. The *asterisk* indicates a significant difference at an alpha-level of $P = 0.005$. *Error bars* represent 1 SEM
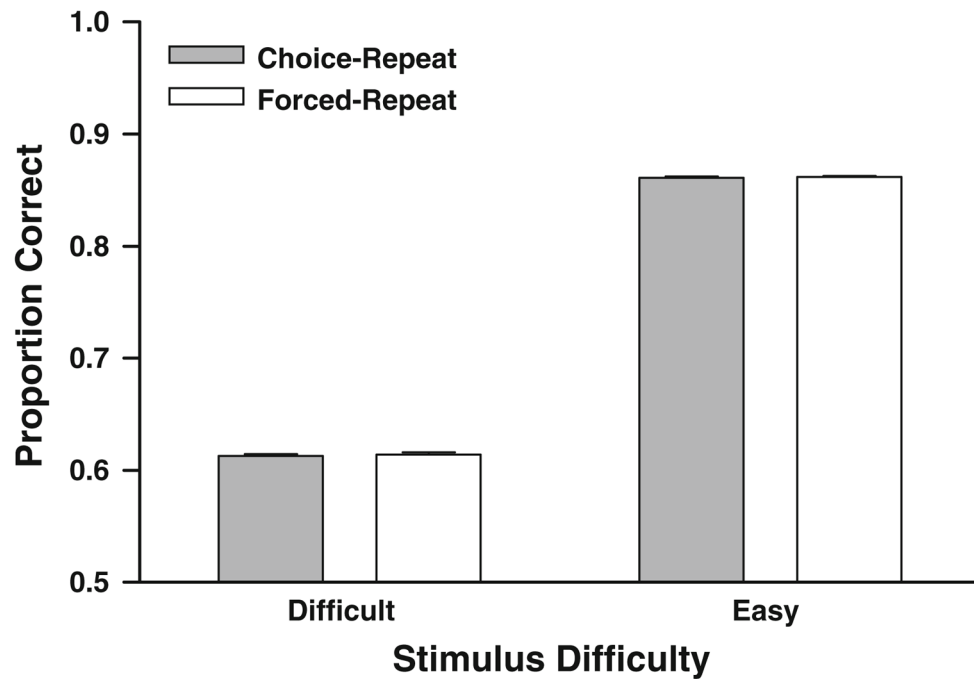
Fig. 4.
Simulated data from the response-strength model. Unlike the empirical data shown in Fig. 3, the simulations show no difference in accuracy between difficult choice-repeat trials and difficult forced-repeat trials. Accuracy on each trial type as a function of stimulus difficulty for simulated data is shown. Difficult stimuli included stimulus Level 31 (i.e., stimulus duration 3.62 s). Easy stimuli included stimulus Level 1 and 11 (i.e., stimulus durations 2.00, 2.44 s). The simulated rate of choosing to repeat the stimulus was 0.578 and 0.730 for easy and difficult conditions, respectively. *Error bars* represent 1 SEM

Table 1

Descriptive statistics for performance in forced-take and choice-take conditions

Condition	N	Mean accuracy (SEM)	
		Difficult stimuli	**Easy stimuli**
Forced-take	7	0.58 (0.016)	0.81 (0.031)
Choice-take	7	0.46 (0.091)	0.84 (0.055)

Table 2

Minimum, maximum, and step-size values

Parameter	Minimum value	Maximum value	Step size
Threshold mean	0.1	1.0	0.1
Threshold SD	0.0	0.0	0.0
SenseS mean	0.005	0.305	0.05
SenseS SD	0.005	0.305	0.1
SenseL mean	0.005	0.305	0.05
SenseL SD	0.005	0.305	0.1
Stimulus SD	0.0	25.0	5.0
Weighted average	0.0	1.0	0.2

The stimulus mean was 1, 11, and 31 in separate simulations. The stimulus range was 1–71, which corresponds to 2.0 and 8.0 s. Easy conditions correspond to durations 2.00 and 2.44 s and stimulus categories 1 and 11 in the simulation. Difficult conditions correspond to the duration of 3.62 s and stimulus category 31 in the simulation

Table 3

Least-squares best fit parameter values

Parameter	Best fit
Threshold mean	0.2
Threshold SD	0.0
SenseS mean	0.255
SenseS SD	0.005
SenseL mean	0.055
SenseL SD	0.205
Stimulus SD	20.0
Weighted average	0.0