To be published in *Behavioral and Brain Sciences* (in press)

© Cambridge University Press 2003

*Below is the copyedited final draft of a BBS target article that has been accepted for publication. This updated preprint has been prepared for formally invited commentators. Please do not write a commentary unless you have been formally invited. (If you are interested in submitting a continuing commentary in the future, please do so after the issue has been published and you are able to read the entire treatment.)*

# The comparative psychology of uncertainty monitoring and metacognition

John David Smith
Department of Psychology
State University of New York at Buffalo
Buffalo, NY 14260
psysmith@acsu.buffalo.edu
http://wings.buffalo.edu/psychology/labs/smithlab/

Wendy Ellen Shields
Department of Psychology
University of Montana
Missoula, MT 59812
wshields@selway.umt.edu
http://psychweb.psy.umt.edu/faculty/shields/shields.html

David Alan Washburn
Department of Psychology
Georgia State University
Atlanta, GA 30303
dwashburn@gsu.edu
http://www.gsu.edu/~wwwpsy/faculty/washburn.htm

**Abstract**: Researchers have begun to explore animals' capacities for uncertainty monitoring and metacognition. This exploration could extend the study of animal self-awareness and establish the relationship of self-awareness to other-awareness. It could sharpen descriptions of metacognition in the human literature and suggest the earliest roots of metacognition in human development. We summarize research on uncertainty monitoring by humans, monkeys, and a dolphin within perceptual and metamemory tasks. We extend phylogenetically the search for metacognitive capacities by considering studies that have tested less cognitively sophisticated species. By using the same uncertainty-monitoring paradigms across species, it should be possible to map the phylogenetic distribution of metacognition and illuminate the emergence of mind. We provide a unifying formal description of animals' performances and examine the optimality of their decisional strategies. Finally, we interpret animals' and humans' nearly identical performances psychologically. Low-level, stimulus-based accounts cannot explain the phenomena. The results suggest granting animals a higher-level decision-making process that involves criterion setting using controlled cognitive processes. This conclusion raises the difficult question of animal consciousness. The results show that animals have functional features of or parallels to human conscious cognition. Remaining questions are whether animals also have the phenomenal features that are the feeling/knowing states of human conscious cognition, and whether the present paradigms can be extended to demonstrate that they do. Thus the comparative study of metacognition potentially grounds the systematic study of animal consciousness.

**Keywords:** cognition; comparative cognition; consciousness; memory monitoring; metacognition; metamemory; self-awareness; uncertainty; uncertainty monitoring

## 1. Introduction

Extensive research on metacognition and uncertainty monitoring has been done to explore humans' capacity to recognize uncertainty and to know when they do not know (Brown et al. 1982; Dunlosky & Nelson 1997; Flavell 1979; Koriat 1993; Metcalfe & Shimamura 1994; Nelson 1992, Nelson & Dunlosky 1991; Nelson & Narens 1990; Reder 1996; Schwartz 1994). Human adults and older human children (hereafter humans) respond adaptively when facing difficult or uncertain situations – they defer response and seek help, hints, or additional information.

Less research has been done to explore the metacognitive capacities of nonhuman animals (hereafter animals – Hampton 2001; Inman & Shettleworth 1999; Shields et al. 1997; Smith et al. 1995; Smith et al. 1997; Smith et al. 1998; Smith & Schull 1989; Teller 1989). We consider the present status and future prospects of this area of comparative psychology. We believe the area would benefit if interested colleagues evaluated existing research and guided future research through their commentaries.

The article proceeds as follows. First, we describe a theoretical framework for metacognition. Second, we discuss potential contributions from a comparative psychology of metacognition. Third, we discuss the requirements for comparative metacognition paradigms that have caused this field to develop slowly. Fourth, we consider existing uncertainty-monitoring results in the domains of perception (Shields et al. 1997; Smith et al. 1995; Smith et al. 1997) and memory (Hampton 2001; Smith et al. 1998). Fifth, we extend phylogenetically the search for metacognitive capacities in animals by considering studies that tested less cognitively sophisticated species (rats and pigeons – Inman & Shettleworth 1999; Smith & Schull 1989; Teller 1989). It is a potentially important fact that these species have not shown evidence of metacognitive capacities, whereas humans, monkeys, and a dolphin have. Sixth, we provide a unifying formal description of performance in the tasks herein described. Seventh, we discuss the appropriate psychological interpretation of animals' performances. Eighth, we consider the nettlesome relationship between these uncertainty-monitoring performances and the declarative consciousness of uncertain mental states.

## 2. A theoretical framework for metacognition

Metacognition is defined to be cognition about cognition. The idea in this field is that some minds contain a cognitive executive that looks in on thought or problem solving to see how it is going and how it might be facilitated (e.g., as when we realize that a paragraph of an article has not been understood and reread it). Nelson and Narens (1990) gave the literature on human metacognition a useful theoretical framework (Figure 1). They theorized that mental activities occur at a meta level and at a lower, object level during cognitive processing. The meta level monitors the processing and determines its progress and prospects. These monitoring functions of the meta level (i.e., the basic metacognitive judgments) are shown at the top of Figure 1. There is an ease-of-learning judgment about whether material will be easy or hard to learn (pepperoni-pizza vs. pepperoni-thesaurus). There is a judgment of learning about the level of learning achieved. There is a feeling-of-knowing judgment about whether information is potentially available in memory or not (e.g., the middle names of William___Clinton vs. Anthony___Blair). There is a judgment about one's confidence in the accuracy of retrieved answers.

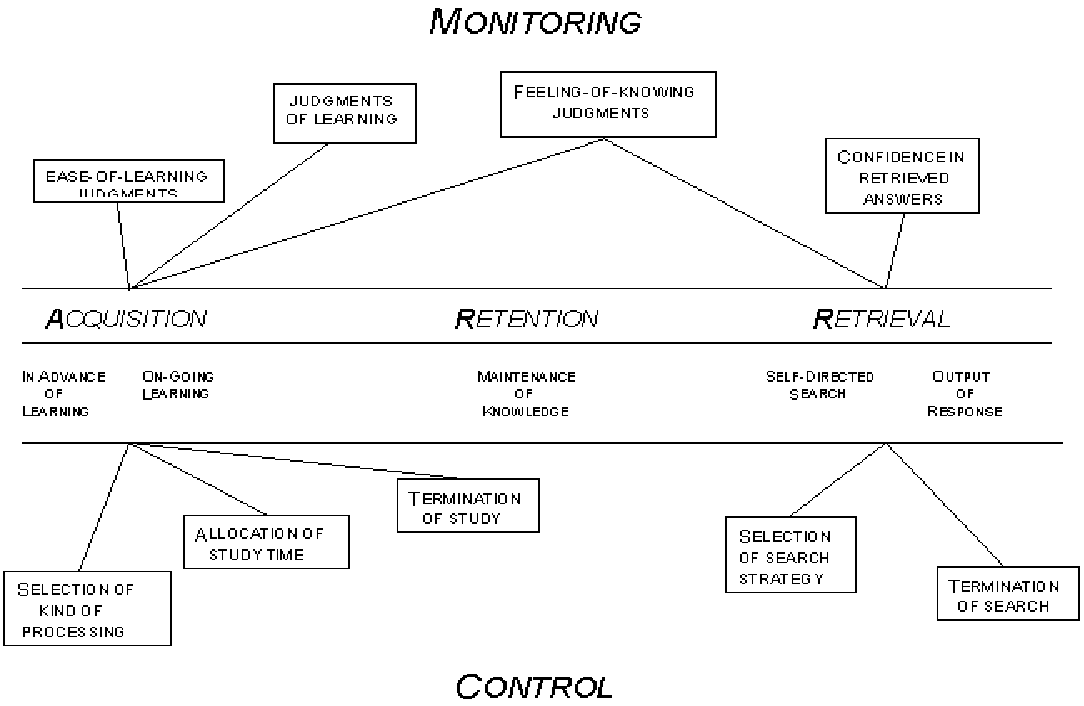


Figure 1. A theoretical framework for research on metacognition, showing examples of process-monitoring capacities above and process-control capacities below (after Nelson & Narens 1990).

The meta level also controls cognitive processing, directing information processing in ways that may be more felicitous. These metacognitive control processes are shown at the bottom of Figure 1. The meta level may select new processes (e.g., elaborative rehearsal instead of maintenance), allocate extra study time to difficult items, terminate studying when sufficient learning is judged to have been achieved (some undergraduates make this judgment too optimistically), select new retrieval strategies when present ones are failing, or abandon retrieval efforts if success seems improbable.

Figure 1 lays out an ambitious potential research program that could show all of these capacities in animals but which has barely begun. Thus far researchers have evaluated whether animals monitor the certain or uncertain status of ongoing perception and memory. No studies have considered whether animals use their uncertainty to adaptively alter the character of information processing. So there remain interesting lines of research to be pursued regarding both the metacognitive control processes and the metacognitive monitoring processes shown in Figure 1.

## 3. Potential contributions of a comparative psychology of metacognition

The comparative study of metacognition potentially illuminates important issues in comparative psychology and cognitive science. First, metacognition is considered one of humans' most sophisticated cognitive capacities and possibly a uniquely human cognitive capacity. It is an important question whether this capacity extends to other species. Second, the comparative study of metacognition would expand the study of animal self-awareness that has depended on the elegant but controversial mirror dye-mark test that assays animals' bodily self-awareness (Gallup 1982; Gallup & Suarez 1986; Parker et al. 1994; Swartz 1997; Swartz et al. 1999). Direct measures of cognitive self-awareness, which may be a different thing from self-recognition (Cheney & Seyfarth 1990, p. 240) would be a useful addition to this area.

Third, comparative metacognition research would contribute to theory of mind research. Theory of mind research asks whether animals know and monitor the other's mental states and states of knowing (Byrne & Whiten 1991; Cheney & Seyfarth 1990; Heyes 1998, and associated commentaries; Whiten & Byrne 1997). The complementary question is whether animals know and monitor their own mental states and states of knowing (Schull & Smith 1992). The relation between these capacities is an important issue in discussions about theory of mind and the evolution of social intelligence. Some theories stress the evolutionary interdependence between self- and other- mental awareness. Perhaps self-awareness evolved in social species to facilitate other-awareness and social intelligence (see Humphrey 1976). Indeed, perhaps self-awareness was a prerequisite for other-awareness. Direct measures of cognitive self-awareness could let theorists explore such possibilities.

Fourth, given the link between humans' metacognition and declarative consciousness (Nelson 1996), the study of animal metacognition would contribute to the study of animal consciousness. Weiskrantz (1986; 1997, ch. 4, pp. 77-99) discussed the possibility of studying animal consciousness through behavioral responses. In his thought experiment, animals had available two discrimination responses and also a commentary key with which to step outside the discrimination and report on the state of their knowledge or perception. The tasks described in this article provide to animals this commentary key in a variety of settings. We agree that these tasks suggest the possibility of exploring animal consciousness systematically. However, we will discuss the difficulty of inferring declarative consciousness from animals' performances in tasks of this kind. We will also discuss the possible theoretical separation between animals' having the functional parallels of humans' conscious cognition and animals' having the subjective, phenomenal feelings of doubt and knowing.

Fifth, because the same uncertainty-monitoring paradigms can be used across several species, metacognitive capacities can be assayed comparably in different phylogenetic groups. This means that it should eventually be possible to draw the map of the phylogenetic distribution of metacognition or cognitive self-awareness. This map might illuminate the emergence of mind.

Sixth, species differences in these capacities might reveal the cognitive mechanisms underlying metacognition. For example, comparing the capacity for uncertainty monitoring in species with and without language could indicate the role of language in metacognition. Correlating uncertainty awareness with the capacities for planning and future-oriented cognition could show the extent to which metacognition is prospective and allied to other executive functions. The comparative psychology of metacognition could also suggest the earliest roots of metacognition in human development and provide the techniques for investigating them. In these and other ways we believe the parallel investigation of animal and human metacognition could produce a constructive synergy.

## 4. Requirements for a comparative uncertainty paradigm

There are two basic requirements for creating a comparative uncertainty paradigm. One is to create perceptual or cognitive difficulty for the animal in order to stir up something like an uncertainty state. The other is to provide a behavioral (i.e., nonverbal) response that lets the animal comment on or cope adaptively with that state. (This second requirement explains why comparative metacognition research began slowly. The typical human paradigms did not suit animals, for their phenomena relied heavily on verbal self-reports about feelings of knowing, judgments of learning, tip-of-the-tongue experiences, and so forth – Brown 1991; Hart 1965; Smith et al. 1991).

These two requirements can be illustrated by considering psychophysical procedures. These procedures are highly developed for creating difficulty for animals. They narrow the contrast between alternative stimulus input classes and force observers to make difficult discriminations at their perceptual thresholds (Au & Moore 1990; Blough 1958; Schusterman & Barrett 1975; Yunker & Herman 1974). They always cause the animal difficulty. The important question for metacognition research is whether animals sense this difficulty and could respond adaptively to it – that is, whether psychophysical procedures generate useable uncertainty states in animals.

Causing the animal difficulty is not sufficient for answering this question. Illustrating this point, Figure 2 shows the performance of a dolphin (*Tursiops truncatus*) near threshold in a discrimination between 2,100-Hz tones and tones at any lower frequency (Smith et al. 1995). The animal made one response (Low) correctly to trials below about 2,075 Hz and the other response (High) correctly to many 2,100-Hz trials. The trials surrounding 2,085 Hz, his known threshold relative to 2,100 Hz (Herman & Arbeit 1972), produced near-chance performance. This task causes the animal the difficulty it should. It might be creating the uncertainty states the comparative metacognition researcher seeks to study. But it cannot show whether the animal senses the difficulty or could cope with the uncertainty. These capacities are hidden by allowing only two responses that map to the two input classes (2,100-Hz tones and lower tones) and by denying the animal any way to comment on uncertainty or respond adaptively to it.

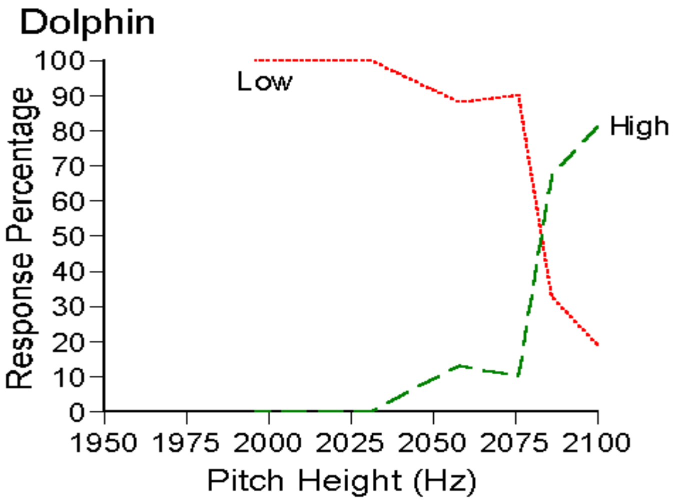


Figure 2. Frequency discrimination performance by a bottlenosed dolphin (*Tursiops truncatus*) in the procedure of Smith et al. (1995). The horizontal axis indicates the frequency (Hz) of the tone. The high response was correct for tones at exactly 2,100 Hz, and these trials are represented by the rightmost data point for each curve. All other tones deserved the low response. The percentages of trials ending with the high response (dashed line) or low response (dotted line) are shown.

It is possible that animals' ancillary behaviors would betray their uncertainty or conflict on threshold trials. These hesitations and waverings do sometimes occur, and they made behaviorists uncomfortable because they suggested that animals might be in mental turmoil over difficult trials. Tolman even suggested that these uncertainty behaviors – "lookings or runnings back and forth" (Tolman 1938, p. 27) – could be taken as a behaviorist's definition of animal consciousness (Tolman 1927). As a behaviorist, Tolman sometimes misbehaved. Moreover, 12 years later, Tolman retracted this claim (1938, p. 27), thereby completing his own theoretical "looking or running back and forth" (See also Tolman 1932/1967).

However, it may not be advisable to rely on ancillary behaviors to convey information about the animal's uncertainty processes. These behaviors may not happen, they may be poorly interpretable or measurable, and they may defeat comparative research because animals in different species may react differently when facing uncertainty. For example, unlike uncertain humans, uncertain dolphins do not scratch their heads. Accordingly, the safest course in comparative metacognition research is to give animals of different species the same concrete response that lets them report on or deal with the difficult situation. This is the second requirement of a comparative uncertainty paradigm, to provide animals a third, Uncertain response that lets them cope with difficulty by declining the trials they do not choose to complete. Then a reasonable strategy for the animal is to use the Uncertain response sparingly, when errors on the primary perceptual or cognitive task are judged to be likely. To carry out this strategy, animals must identify, if they can, the occasions on which they are liable to err.

## 5. The uncertain response in human psychophysics

One methodology for studying animals' uncertainty monitoring combines psychophysical procedures that cause animals perceptual difficulty with the uncertainty response that lets them report on and cope with their uncertainty (Smith et al. 1995; Smith et al. 1997). This paradigm has a rich history in human experimental psychology that raises the possibility of studying metacognition comparatively.

Human observers in early psychophysical studies were often allowed to respond Uncertain when they felt unable to assign a stimulus to one of the two primary input classes (Angell 1907; Fernberger 1914, 1930; George 1917; Watson et al. 1973; Woodworth 1938). Some researchers hoped that the level of uncertainty responses would index sensory sensitivity (e.g., Urban 1910), with dull and sharp perceivers needing that response more and less, respectively. However, others questioned this approach (Fernberger 1914; Woodworth 1938, pp. 419-27), noting the special psychological status of the response Uncertain. That response had longer latencies (Angell 1907; George 1917; Woodworth 1938) and special susceptibility to instructions (Brown 1910; Fernberger 1914, 1930; Woodworth 1938). It seemed to be linked to participants' personality or temperament (Angell 1907; Fernberger 1930; Thomson 1920) – a linkage that surely had little to do with studying sensory capability. It seemed to be more reflective and cognitive than the two primary discrimination responses (Angell 1907; Fernberger 1930; George 1917).

Theorists also had a structural concern about uncertainty responses. They believed that these were meta to the primary discrimination and were a comment on the participant's failure to assign a stimulus to one of the primary input classes. George (1917) concluded that the doubtful responses were offenders against the constant attitude required in psychophysics because they introduced "extra-serial" attitudes into a task that depended on intra-serial, sensory attitudes. Boring (1920) suggested that doubt was an attitudinal seducer that took observers away from the series of mental states that are a continuous function of the series of stimuli. Jastrow (1888) and Brown (1910) recommended forcing a primary discrimination response from observers on each trial, while collecting confidence judgments on the side to replace the information provided by uncertainty responses. Confidence ratings are often collected from humans in this way. The catch is that animals have so far not proved able to report their confidence in this way (but see Shields et al. in press).

However, we will see that animals can respond Uncertain. Given that fact, the structural problem early theorists had with this response makes it intriguing regarding a comparative psychology of uncertainty monitoring. The Uncertain response might be meta to animals' primary discriminatory process, too. It might be for them, too, a comment on or a response to indeterminacy and difficulty in the primary discrimination. There is contemporary convergence on this possibility. The psychophysical Uncertain response instantiates the commentary key that Weiskrantz (1986, 1997) considered providing to blindsight animals in a thought experiment. Cowey and Stoerig (1992) proposed a similar Gedanken procedure (see also Cowey & Stoerig 1995). Their idea was to train a light no-light discrimination, then use psychophysical procedures to bring the animal to threshold where it would be only 50% correct. Critically, though, the animal would also have a third lever reinforced on a 75% schedule. Then the adaptive animal could report on its state of not knowing whether a light was seen by choosing the third lever selectively on threshold trials.

These converging ideas made it clear that the psychophysical uncertainty paradigm was a strong starting place for the comparative study of uncertainty monitoring. We discuss now the results when humans, monkeys, and a dolphin were placed into difficult and uncertain perceptual situations and forced to make judgments at threshold, but were also allowed to respond Uncertain.

As we discuss these results, readers will naturally consider both metacognitive interpretations of performance and possible lower-level interpretations of performance. For example, there might be inadvertent cueing of the animal (e.g., by a dolphin trainer) to respond Uncertain on difficult trials. Or, the animal might respond Uncertain to avoid aversive, error-producing stimuli instead of doing so to cope adaptively with uncertainty. Additional research in this field (Sections 8-10) has addressed some alternative accounts. Alternative accounts are also discussed in Section 14 on the problem of psychological interpretation. Section 14.2 specifically discusses the role that low-level, associative explanations of behavior have in explaining comparative data.

## 6. Uncertain responses by humans and monkeys in a psychophysical density discrimination

To begin the comparative study of uncertainty monitoring, Smith et al. (1997) placed humans and rhesus monkeys (*Macaca mulatta*) in a visual density-discrimination task.

Participants used a joystick to move a cursor to one of three objects on a computer screen (Figure 3a). The dense response (choosing the box) was correct if the box contained exactly 2,950 illuminated pixels. The sparse response (choosing the S) was correct if the box had any fewer pixels. The Uncertain response (choosing the star) allowed participants to decline the trial and move into a new, guaranteed-win one. Initially, participants were stabilized on an easy discrimination involving 2,950- and 450-pixel boxes. Then the discrimination's difficulty was gradually increased by making the sparse boxes denser until performance faltered at about 2,950 vs. 2,600 pixels. At mature performance, trial difficulty was continuously adjusted based on participants' performance within a session to maintain difficulty at a constant, high level.

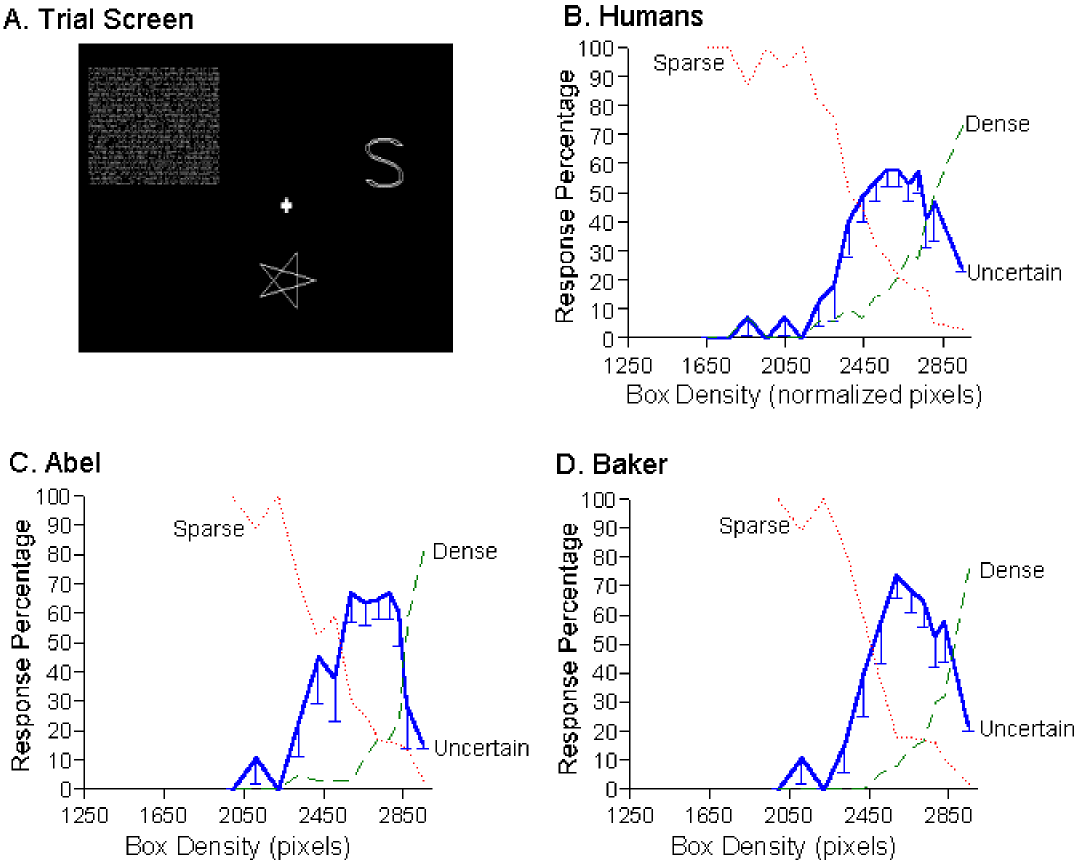


Figure 3. A. The screen from a trial in the dense-sparse discrimination of Smith et al. (1997). B. The performance of seven humans in the dense-sparse task. The dense response was correct for boxes with exactly 2,950 pixels – these trials are represented by the rightmost data point for each curve. All other boxes deserved the sparse response. To equate discrimination performance across participants, the data have been normalized to place each participant's discrimination crossover at a pixel density of about 2,700. The horizontal axis indicates the normalized pixel density of the box. The solid line represents the percentage of trials receiving the Uncertain response at each density level. The error bars show the lower 95% confidence limits. These were calculated (Hays 1981, pp. 224-26) using the total Uncertain responses as a proportion of total trials at each density level (summing across humans who completed one session each). The percentages of trials

ending with the dense response (dashed line) or sparse response (dotted line) are also shown. C. The performance of Monkey Abel in the dense-sparse discrimination depicted in the same way (here the error bars were calculated using the total Uncertain responses as a proportion of total trials at each density level summing across multiple sessions by the animal). D. The performance of Monkey Baker in the dense-sparse discrimination.

Figure 3b shows the performance of seven humans. Sparse responses predominated on sparser trials; dense responses predominated on true dense trials and the most difficult sparse trials. The primary discrimination was performed at chance where these two response curves cross. Humans responded Uncertain most in the region of uncertainty around this perceptual threshold. Humans knew when they were at risk for error in the primary discrimination and declined those trials selectively and adaptively.

Humans' post-experimental reports constructively corroborated this primary behavioral evidence of uncertainty responding. They said that their sparse and dense responses were cued by the objective stimulus conditions (i.e., sparsity or density) on a trial. In contrast, they said that their Uncertain responses were prompted by personal feelings of uncertainty, doubt, and of not knowing the correct answer in the discrimination (I was uncertain; I didn't know or couldn't tell). This suggests that, for humans, the Uncertain response may reveal not only metacognitive monitoring but also a reflexive awareness of the self as cognitive monitor. Given their construal of the Uncertain response, some humans did not even like using it, for they felt it was cheating or a cop-out. Humans have given these same introspections about Uncertain responses – that is, that they are "sort of an admission of weakness" (Fernberger 1930, p. 210) – for almost 100 years. In related recent work we have found that males are especially overconfident in a task of this kind and tend to think they know even when they do not (Washburn et al. 2001). Apparently, males are unlikely to stop and ask for directions even within psychophysical discriminations.

These data replicated the phenomenon and the phenomenology of the Uncertain response in human psychophysics. They confirmed that this response is a comment by humans on the failure of the primary discriminatory process, or a no-confidence vote on that process. Humans' metacognitive performance in this task thus provided a good comparative target to which monkeys' performance could be referred.

Two 9-year-old rhesus monkeys participated in the same task. These joystick-trained monkeys were tested using the Language Research Center's Computerized Test System (LRC-CTS; Washburn & Rumbaugh 1992). They received food pellets or 19-s timeouts, respectively, for correct or incorrect responses.

Both monkeys performed like the humans did, with Uncertain responses focused on the discrimination's crossover (Figures 3c,d). The monkeys declined somewhat more trials than the humans did, possibly because the monkeys (but not the humans) were working for food rewards, or possibly because the humans (but not the monkeys) had scruples about using this response. Monkeys, like humans, correctly assessed when they were at risk for error in the primary discrimination and declined those trials preemptively. There is clearly a strong analogy between the use of the psychophysical uncertainty response by the two species. In fact, Figure 3 presents one of the strongest existing matches between human and animal performance in the comparative literature.

A related experiment strengthened this analogy by demonstrating parallel individual differences within the two species in the use of the Uncertain response. In this experiment, instead of dynamically titrating perceivers' thresholds, we continuously offered perceivers a wide range of densities from easy sparses to easy denses. This experiment resembled others that have collected doubtful judgments from humans (e.g., Woodworth 1938).

Figures 4a,b, respectively, show the results from eight humans and from one particular human. The two discrimination responses (sparse and dense) were used in the same way. The difference is that humans generally responded Uncertain to the difficult, indeterminate trials near the discrimination's breakpoint. The one human did not. Figures 4c,d, respectively, show the performance of Monkeys Baker and Abel. Both animals used the two discrimination responses in the same way, performing better for stimuli farther from the task's midpoint. Baker selectively declined the indeterminate trials near the discrimination's breakpoint that he would most probably lose; Abel did not. Thus, under the same circumstances that caused a human not to respond Uncertain in this task, Abel chose not to, either.

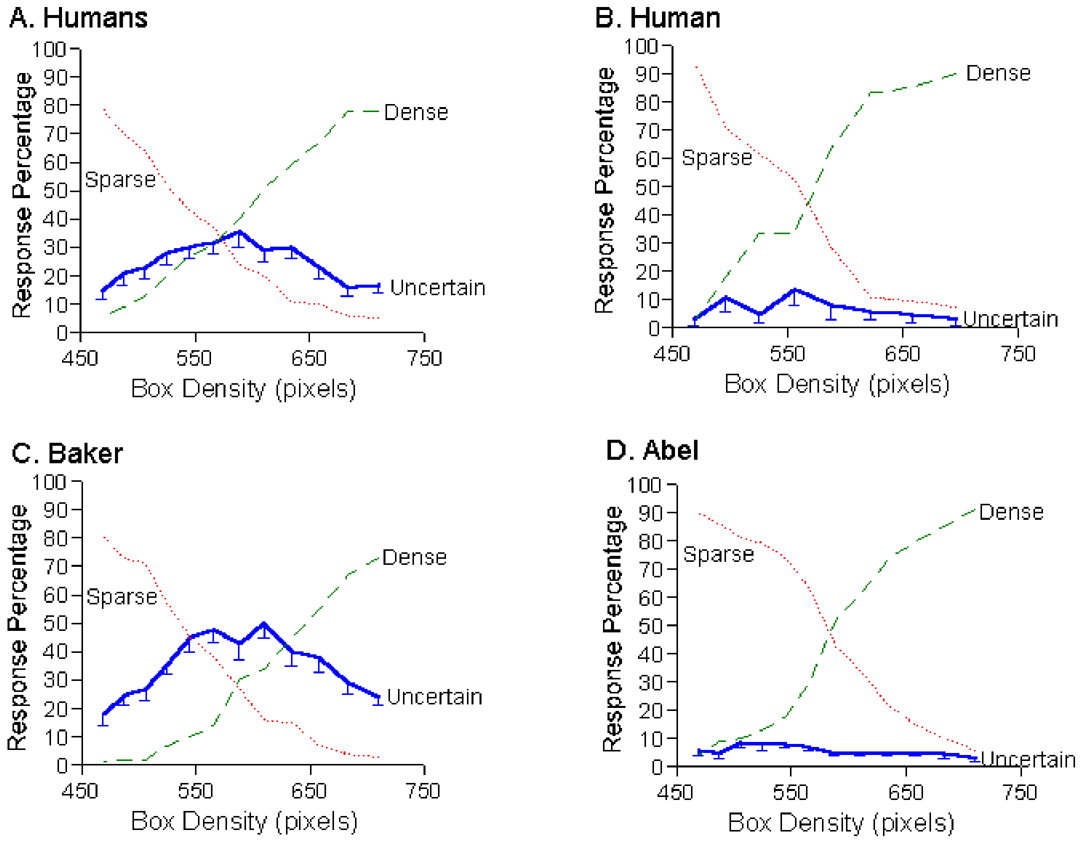


Figure 4. A. The performance of eight humans in the broad-spectrum dense-sparse discrimination of Smith et al. (1997). The upper half and lower half of a range of densities deserved the dense response and sparse response, respectively. The percentage of trials receiving the Uncertain response at each density are shown by the solid line. The error bars show the lower 95% confidence limits. The percentages of trials ending with the dense response (dashed line) or sparse response (dotted line) are also shown. B. The performance of one human in the broad-spectrum dense-sparse discrimination. C. The performance of Monkey Baker in this discrimination. D. The performance of Monkey Abel in this discrimination.

All humans and animals in both dense-sparse tasks just described used the primary discrimination responses (dense and sparse) in the same way. This supports the idea that the two primary responses in a discrimination task are functionally highly stable for being perceptually anchored by and mapped to the two stimulus input classes. Across experiments, amongst humans and between monkeys, only the use of the Uncertain response showed strong differences. The early psychophysicists, observing this peculiar changeability of the Uncertain response in the hands of humans, concluded that the Uncertain response was not perceptually anchored to a stimulus input class, that it was about the failure of assignment of stimuli to an input class, and that it was related to uncertainty, extra-serial attitudes, and decisional temperaments (Angell 1907; Fernberger 1930; Thomson 1920). It is an interesting fact that monkeys' use of the Uncertain response shows this peculiar changeability, too.

## 7. Dolphin uncertainty responses in an auditory discrimination

Smith et al. (1995) evaluated the uncertainty-monitoring capacities of another cognitively sophisticated species by placing a dolphin in an auditory discrimination task. Pressing the high or low paddles, respectively, was correct for 2,100-Hz tones or tones of any lower frequency. The Uncertain paddle advanced the animal into an easy, low-pitched trial that was rewarded when completed with the low response. Initially, the animal was stabilized on an easy discrimination between 2,100 Hz and 1,200 Hz. Then the discrimination's difficulty was increased by raising the pitch of the below-2,100 Hz trials until the dolphin was struggling to distinguish trials of 2,100 Hz and 2,085 Hz. At mature performance, trial difficulty was adjusted based on the dolphin's performance to sustain the level of difficulty.

Figure 2 showed the dolphin's two-response performance with the Uncertain response disallowed. Low and high responses mapped to below-2,100 Hz and 2,100 Hz tones, respectively, with these response curves crossing (signifying chance performance) at the dolphin's threshold. The crucial question, not illuminated by Figure 2, was how the animal would behave at threshold when allowed to respond Uncertain.

Figure 5a answers this question. The dolphin's primary discrimination performance was the same, but now he used the Uncertain response for the difficult trials surrounding his discrimination threshold. Five humans performed similarly (Figure 5b). Both species used the Uncertain response less in this auditory discrimination than did monkeys and humans in the density discrimination. That is, the perceivers in the auditory task had a narrower Interval of Uncertainty (Woodworth 1938). Probably the density continuum has a wider region of subjective indeterminacy because the placement of the dots in the boxes has a random element that can cloud one's perception of denseness and sparseness. Probably the pitch continuum has a narrower region of subjective indeterminacy because the pitch continuum is unidimensional and pure. It is interesting that both human and animal observers sense similarly and accurately the breadth of the zone of subjective indeterminacy and use the Uncertain response appropriately to each sensory domain.

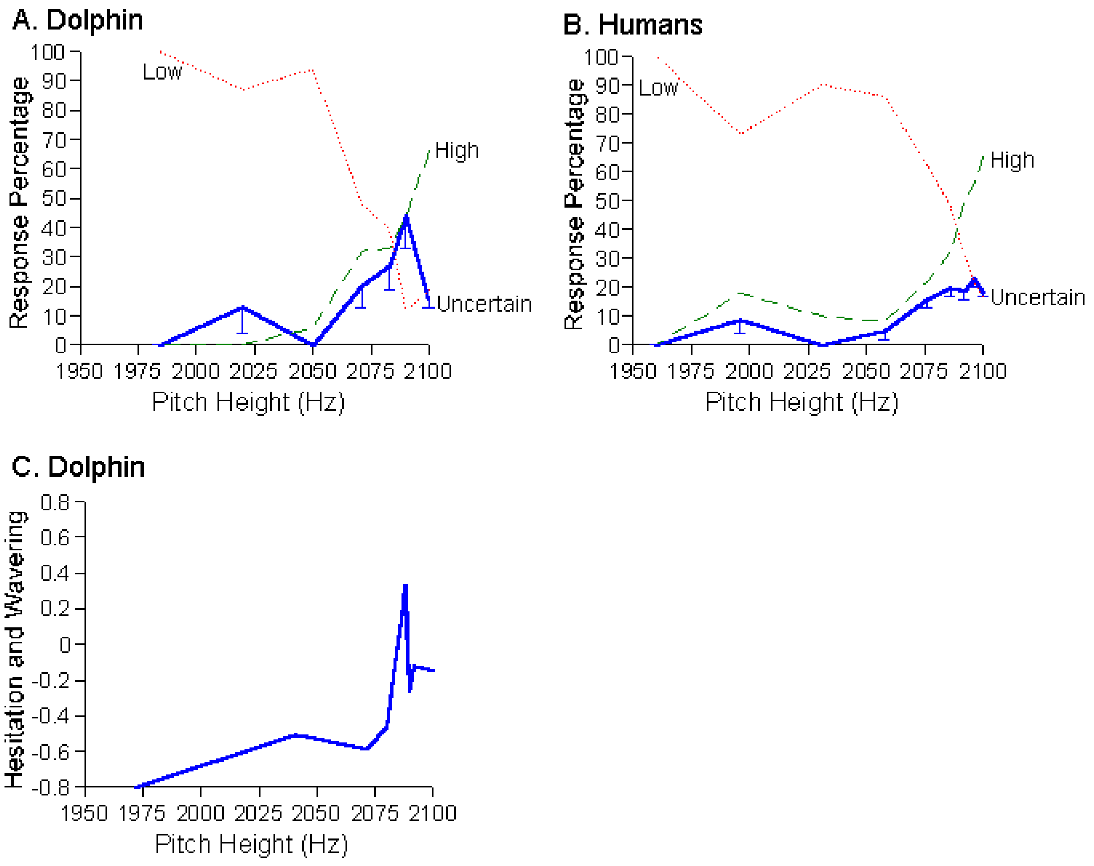


Figure 5. A. Performance by a dolphin in the auditory discrimination of Smith et al. (1995) when the Uncertain response was also available. The horizontal axis indicates the frequency (Hz) of the trial. The high response was correct for tones at 2,100 Hz – these trials are represented by the rightmost data point for each curve. All other tones deserved the low response. The solid line represents the percentage of trials receiving the Uncertain response at each difficulty level. The error bars show the lower 95% confidence limits. The percentages of trials ending with the high response (dashed line) or low response (dotted line) are also shown. B. The performance of five humans in a similar auditory discrimination. C. The dolphin's weighted overall Factor 1 behavior (hesitancy, slowing, wavering) for tones of different frequencies (Hz).

Humans again attributed their use of the two primary discrimination responses (high and low) to the prevailing stimulus conditions (i.e., 2,100 Hz tones and lower tones). They attributed their Uncertain responses, as the early psychophysical observers did, to their states of doubt and uncertainty. Though the dolphin said nothing, an interesting additional result was that his own brand of uncertainty behaviors attended his Uncertain responses near threshold. He sometimes slowed approaching the response paddles, or wavered amongst them, or swam toward them with an open mouth, or while sweeping his head from side to side or opening and closing his mouth rhythmically. To formalize these observations, four raters judged the intensity of these behaviors during the trials in four video-taped sessions. Then factor analysis was used to discern the latent structure behind the correlations among these variables. The strongest behavioral factor was clearly allied to hesitation and wavering by the animal. Figure 5c shows the overall intensity of these Factor 1 behaviors (based on a factor-scoring procedure) for trials at different pitch levels. These behaviors peaked at 2,087 Hz and were distributed like the Uncertain response (Figure 5a). These behaviors are the "lookings or runnings back and forth" that Tolman (1938, p. 27) thought might operationalize animal consciousness. They are also intuitive symptoms of uncertainty states in the animal. Thus these ancillary behaviors reinforce an uncertainty interpretation of the animal's Uncertain responses. However, the Uncertain response is more easily measured and compared across situations and species than are hesitation and wavering.

## 8. Uncertain responses by monkeys and humans in a same-different task

The experiments just summarized focused on stimulus qualities (e.g., 2,100 Hz or lower; true dense or sparser). Those experiments leave open the possibility (discussed in Section 14) that these Uncertain responses by animals fell under the associative control of stimulus cues rather than under the metacognitive control of uncertainty states. This possibility led Shields et al. (1997) to ask whether monkeys could recruit adaptive uncertain responses when pushed to their psychophysical limit in a same-different (SD) task. The SD task, if constructed correctly, requires a relational judgment, and an abstraction beyond the current absolute stimulus qualities, especially when sameness and difference must be judged amid variable stimulus contexts. Probably the SD task requires an additional processing step in which the relevant qualities of the two stimuli are compared to or subtracted from one another (a differencing strategy – see MacMillan & Creelman 1991). Then, as the difference resulting is near zero or larger than zero, a judgment of same or different is made. This information-processing description grounds the prominent idea that relational concepts are cognitively derived, sophisticated, and phylogenetically restricted. Animals often have special difficulty with relational judgments (Carter & Werner 1978; Premack 1978; see also Cumming & Berryman 1961; Farthing & Opuda 1974; Fujita 1982; Holmes 1979) and require clever training procedures to learn them (Wright et al. 1990). Moreover, relational concepts can be fragile when placed into opposition with absolute stimulus qualities (Premack 1978). For these reasons relational learning heads standard typologies of conceptual sophistication (Herrnstein 1990). For these same reasons it is interesting to know whether animals can make adaptive uncertain responses in this abstract setting.

Accordingly, Shields et al. (1997) gave rhesus monkeys a same-different density discrimination. On each trial two rectangles filled randomly with lit pixels were shown (Figure 6a). As the two densities shown on a trial were the same or different, animals were to make the same response (moving the cursor to the rectangles), or the different response (moving the cursor to the D). To cause animals serious difficulty, the same-different task was psychophysically scaled. That is, the size of the stimulus disparity on different trials was adjusted dynamically to challenge constantly participants' ability to discriminate same from different. In addition, same and different trials at several absolute stimulus levels were intermixed to ensure a true relational performance.

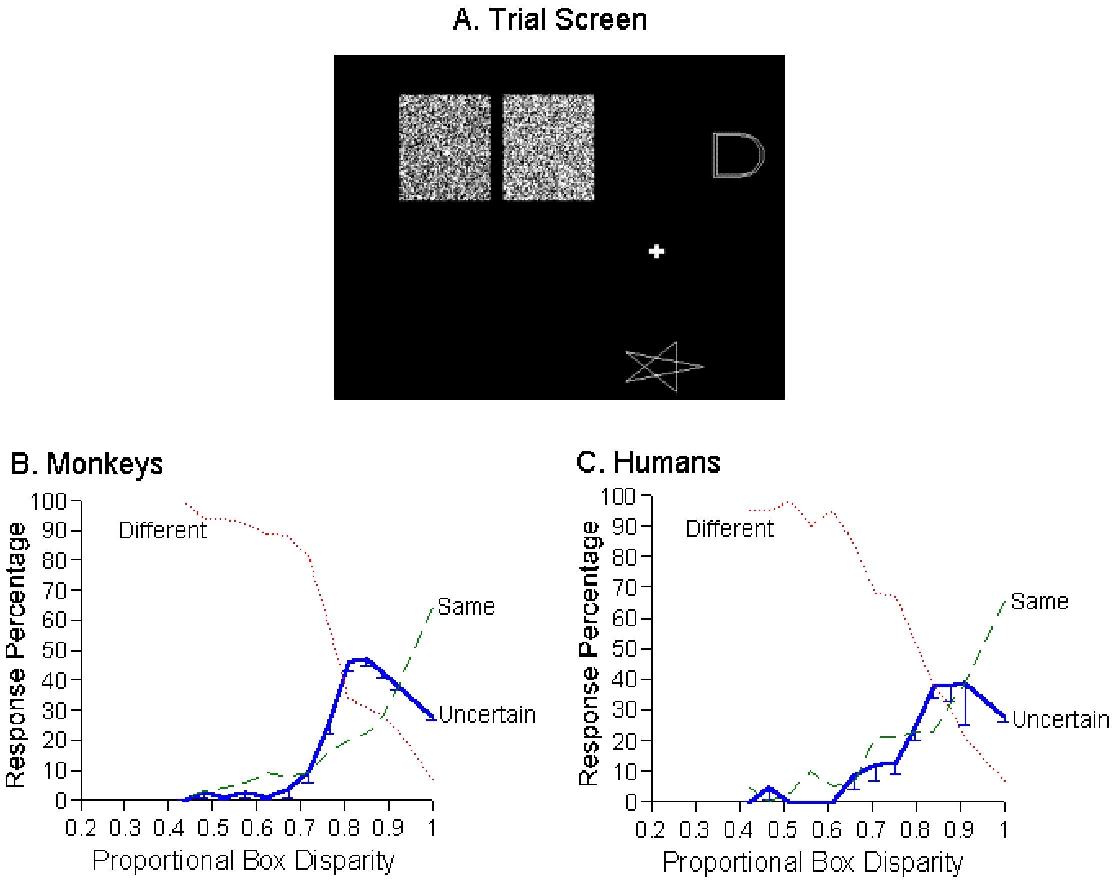


Figure 6. A. The screen from a different trial of the same-different task of Shields et al. (1997). B. Performance by two monkeys in the same-different task. The horizontal axis gives the ratio between the densities of the comparison box and the standard box for trials of different disparities. The same response was correct for trials at a proportional box disparity of 1.0, and these trials are represented by the rightmost data point for each curve. All other trials deserved the different response. The solid line represents the percentage of trials receiving the Uncertain response at each density ratio. The error bars show the lower 95% confidence limits. The percentages of trials ending with the same response (dashed line) or different response (dotted line) are also shown. C. Performance by six humans in the same-different task.

The crucial question is whether monkeys can decline trials that present indeterminate stimulus disparities. In fact they were undeterred by either the difficulty or abstractness of the task. The two monkeys (Figure 6b) used the Uncertain response in just the way that six humans did (Figure 6c). The animal and human performance profiles correlate at *r* = 0.98. Shields et al. (1997) even reserved some absolute density levels for transfer tasks and demonstrated that the animals were showing a true relational performance that was independent of the absolute dense and sparse stimuli that carry the relation.

## 9. Smith et al.'s (1998) comparative studies of memory monitoring

Memory tasks have been a sharp focus in studies of human metacognition. For example, humans can be asked to judge whether they can complete phrases like "The physicist Albert_____" or "The philosopher Albert_____". Comparative research has also asked whether animals can monitor their memory and respond adaptively when the state of their memory does not justify completing a memory test.

Smith et al.'s (1998) exploration of this capacity relied on the predictable changes in memory performance that occur across the serial positions of a memory list. The experimenter can know which items cause difficulty (and perhaps uncertainty) for the animal, and can ask whether animals use the Uncertain response selectively for these difficult memory trials. Smith et al. adopted the serial-probe recognition (SPR) task that has been a staple in comparative memory research (Castro & Larsen 1992; Roberts & Kraemer 1981; Sands & Wright 1980; Wright et al. 1985). In this procedure one presents a "list" of items sequentially followed by a probe. The participant makes a there or not there response as the probe is judged to have been in the list or not. In elegant parametric research using this task, Wright et al. (1985) specified conditions that lead monkeys to show both primacy and recency effects. Our procedures followed on theirs to produce serial-position curves that had this classic shape. But we also gave animals an Uncertain response that let them decline any memory tests of their choosing.

Figure 7a shows that monkeys did decline trials selectively when the middle, more difficult list positions were probed. The tendency to respond Uncertain was the mirror image of memory performance when the animal chose to complete the memory test. Figure 7b shows that the same was true for 10 humans under similar conditions (though these conditions were not suitable for humans to show a strong primacy effect). The similarity in the two graphs is especially interesting because humans were expressly instructed to use the Uncertain response as a report on memory indeterminacy. Monkeys are behaving like humans. Humans are declining memory tests when they feel uncertain.

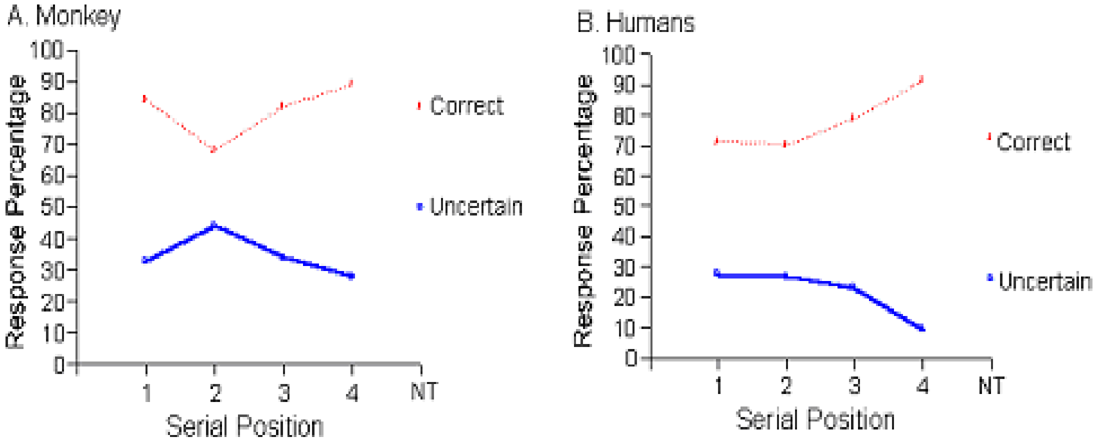


Figure 7. A. Serial probe recognition (SPR) performance by Monkey Baker in the task of Smith et al. (1998). NT denotes not there trials. The serial position (1-4) of the probe in the list of pictures is also given along the X-axis for the probes on there trials. The percentage of total trials that received the Uncertain response is shown (bold line). The percentage correct (of the trials on which the memory test was accepted) is also shown (dotted line). B. Performance by 10 humans in a similar SPR task used by Smith et al.

An additional prediction follows from an uncertainty-monitoring interpretation of monkeys' performance. Animals should perform better when they choose to complete the memory test than when they are forced to do so. This prediction follows because the monkey should accept memory tests when he monitors quite strong or quite weak traces that can correctly be given there or not there responses. In contrast, when one forces the animal to complete all memory tests, his overall performance will be lowered by adding in the poor performance on the memory tests that he would have declined because he monitored indeterminate traces on those trials.

To evaluate this important prediction, Smith et al. (1998) also ran Monkey Baker without the Uncertain response available. Under these conditions, Baker was 72%, 51%, 67%, and 75% correct when serial positions 1 to 4 were probed. His performance was 12%, 17%, 15%, and 14% higher, respectively, when he had the Uncertain response available but chose to complete the memory test. Baker increased his rewards per minute by 18% by using the Uncertain response adaptively to avoid errors when he monitored an indeterminately available memory. Baker was clearly sensing something real about his memory that was rationally attended to in deciding to accept or decline memory tests. This prediction – for a performance advantage on chosen trials over forced trials – figures prominently in a variety of related studies, the next of which we consider now.

## 10. Hampton's (2001) comparative studies of memory monitoring

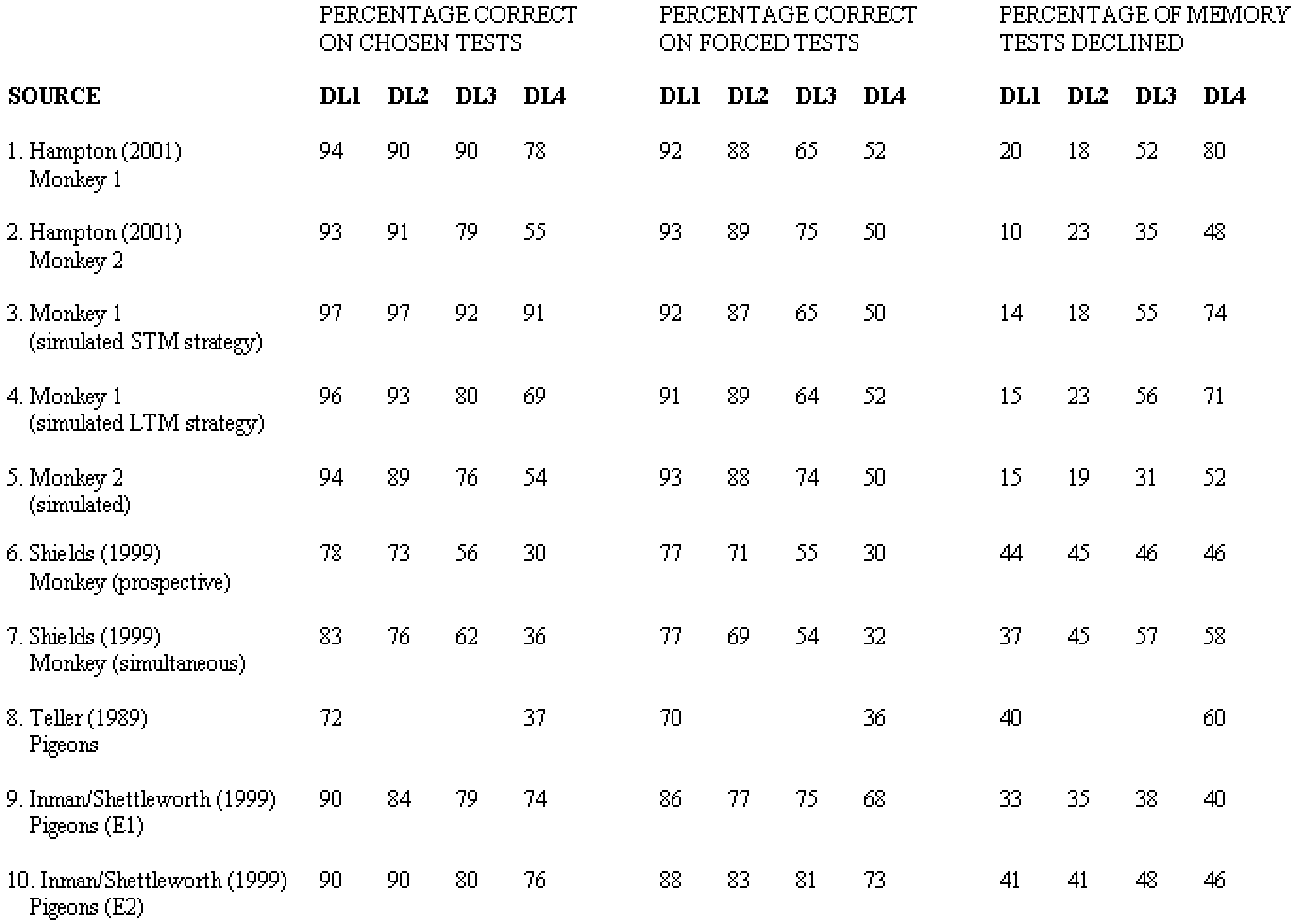
Hampton also asked whether monkeys can use the monitored strengths of memory traces in deciding whether to accept or decline memory tests. Hampton's exploration of this capacity took advantage of the fact that memory performance declines predictably during a forgetting interval. Thus the experimenter can know which items cause difficulty (and perhaps uncertainty) for animals, and can ask whether animals use the Uncertain or Decline response selectively for these difficult memory tests. Hampton adopted the delayed matching-to-sample (DMTS) task that has also been a staple in comparative memory research (Elliott & Dolan 1999; Etkin & D'Amato 1969; Herman 1975). In this procedure one presents a sample that then must be chosen after the forgetting interval from among two or more alternatives.

The critical point in Hampton's procedure came at the end of the forgetting interval on each trial. Then, on 67% of trials, the animal chose, by making one or another discriminative response, to accept the memory test (a four-alternative, forced-choice test in which the sample was presented along with three foils) or to decline it. Accepted memory tests led to either a preferred food reward or a timeout for correct and incorrect responses, respectively. Declined memory tests led to a non-preferred food reward with no risk of timeout. The ideal strategy was to consult the strength or availability of the sample's trace in memory at the end of the forgetting interval, and accept the trial if that strength exceeded some criterion value that would probably support correct recognition in the memory test. By this strategy, the animal would sensibly decline tests when the trace fell below criterion because that weak trace might be unrecognizable amidst the foils at test. On the remaining 33% of the trials, the animal was forced to complete the memory test, just as Monkey Baker was when Smith et al. (1998) ran him without the Uncertain response available. Note that by interspersing chosen and forced memory tests, Hampton was able to simultaneously monitor the animals' performance on these two kinds of trials. This is an innovative approach that both Teller (1989) and Inman and Shettleworth (1999) developed independently. Hampton predicted that forced memory tests would produce poorer performance than chosen memory tests for the reason already given.

Hampton's Experiment 3 is especially interesting and critical. The forgetting interval was varied in several steps from about 15 s to more than 100 s. Hampton made two predictions that would follow if animals were accepting or declining memory tests based on metamemory. First, they should decline more trials as the forgetting interval increased, because the sample's trace would grow less available with time and would more often fall below the criterion level of strength or availability. Second, animals should show a stronger advantage in performance for chosen memory tests over forced memory tests as the forgetting interval increased. This would occur because the strong traces after short delays would support near-ceiling performance on chosen and forced trials, whereas the weaker traces after long delays would be more variable and, if the animal monitored the strength of these traces, would show more clearly the adaptive value of the metacognitive strategy.

One monkey confirmed both predictions (Table 1, Row 1), showing a perfect metamemory data pattern. In the table, DL1 to DL4 denote four increasing difficulty levels, which here refer to longer forgetting intervals. (Hampton did not fully report performance for a fifth, longest delay. The performance estimates in that condition were unstable because nearly all trials were declined. So we did not include this fifth delay condition here.) More memory tests were declined at longer delays. There was better performance on chosen than on forced memory tests, especially at long delays. The second monkey (Table 1, Row 2) presented a different data pattern. This animal did decline more memory tests at longer delays. Alone, though, this result might simply reflect that the animal associated long delays with memory-test errors and timeouts, motivating decline responses after long delays. This animal hardly showed the other crucial component of the metamemory data pattern (performing better on chosen than on forced trials).

Table 1. *Percentage of chosen and forced memory tests answered correctly, and percentage of memory tests declined, for four levels of trial difficulty (DL1 to DL4) in previous studies*



| | PERCENTAGE CORRECT ON CHOSEN TESTS | | | | PERCENTAGE CORRECT ON FORCED TESTS | | | | PERCENTAGE OF MEMORY TESTS DECLINED | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SOURCE** | **DL1** | **DL2** | **DL3** | **DL4** | **DL1** | **DL2** | **DL3** | **DL4** | **DL1** | **DL2** | **DL3** | **DL4** |
| 1. Hampton (2001)<br>Monkey 1 | 94 | 90 | 90 | 78 | 92 | 88 | 65 | 52 | 20 | 18 | 52 | 80 |
| 2. Hampton (2001)<br>Monkey 2 | 93 | 91 | 79 | 55 | 93 | 89 | 75 | 50 | 10 | 23 | 35 | 48 |
| 3. Monkey 1<br>(simulated STM strategy) | 97 | 97 | 92 | 91 | 92 | 87 | 65 | 50 | 14 | 18 | 55 | 74 |
| 4. Monkey 1<br>(simulated LTM strategy) | 96 | 93 | 80 | 69 | 91 | 89 | 64 | 52 | 15 | 23 | 56 | 71 |
| 5. Monkey 2<br>(simulated) | 94 | 89 | 76 | 54 | 93 | 88 | 74 | 50 | 15 | 19 | 31 | 52 |
| 6. Shields (1999)<br>Monkey (prospective) | 78 | 73 | 56 | 30 | 77 | 71 | 55 | 30 | 44 | 45 | 46 | 46 |
| 7. Shields (1999)<br>Monkey (simultaneous) | 83 | 76 | 62 | 36 | 77 | 69 | 54 | 32 | 37 | 45 | 57 | 58 |
| 8. Teller (1989)<br>Pigeons | 72 | | | 37 | 70 | | | 36 | 40 | | | 60 |
| 9. Inman/Shettleworth (1999)<br>Pigeons (E1) | 90 | 84 | 79 | 74 | 86 | 77 | 75 | 68 | 33 | 35 | 38 | 40 |
| 10. Inman/Shettleworth (1999)<br>Pigeons (E2) | 90 | 90 | 80 | 76 | 88 | 83 | 81 | 73 | 41 | 41 | 48 | 46 |

The metamemory studies of Smith et al. and Hampton are similar but have an important difference and complementarity. The difference is that Hampton's monkeys needed to place only one criterion along the continuum of trace-strength impressions. They decided only whether to accept the memory test or decline it. Smith et al.'s monkeys needed to place two criterion lines because if the trace were really available or unavailable they needed to respond there or not there, respectively, so that they only declined the memory test for indeterminate traces. We return to the question of how the single criterion line in Hampton's task relates to the criterion lines in Smith et al.'s task.

The complementarity is that Smith et al. and Hampton followed different lines of research in the human metacognition literature. Smith et al. asked their animals for the decline-accept decision with the memory probe present. The animal had to consider how available that probe's trace was in the context of the memory list – that is, whether it was easily available as a member of the list, unavailable, or indeterminate. Hampton asked his animals for the decline-accept decision with the memory probe absent. The animal had to consider how active relevant traces were in memory and whether they were active or accessible enough to suggest accepting the memory test. Hampton's procedure is a strong one, because the animal was not shown which memory location should be evaluated in making the decline-accept decision. (Inman & Shettleworth 1999, also used this procedure in their Experiment 2 that is discussed in sect. 12.3). The limitation on his method is that only four pictures were ever relevant at a time, so the animal only needed at most to consult the availability of all four traces. Smith et al.'s procedure is a strong one because it is known that re-presenting the material to be judged for the quality of memory remakes those traces active and available and makes metacognition judgments more difficult (Dunlosky & Nelson 1992, 1997). The limitation to this approach is that Smith et al. showed their animals which memory location to monitor for a contextual list memory. In the section on formal approaches we will see that Smith et al.'s and Hampton's procedures are analyzable using the same formal assumptions, suggesting that both studies tapped the same memory-monitoring capacity in monkeys. However, as we will see next, neither procedure meets the full challenge posed by the human paradigms. We describe now an experiment that came closer to doing so, though the monkeys did not meet the challenge because they did not express a metacognitive capacity within it.

## 11. Shields's (1999) comparative studies of memory monitoring

Shields (1999) undertook the comparative study of memory monitoring that is closest to the human paradigms. This study suggests limits on monkeys' metamemory and indicates lines of future research. Human metamemory experiments frequently ask participants to make feeling-of-knowing judgments about linked information that is not presented, so that the query prompts a search for the availability of information at some other memory location (e.g., ______Putin). In contrast, both Smith et al. (1998) and Hampton (2001) asked directly about the availability of memory material that itself had just been presented (e.g., Vladimir – 30 s forgetting interval – Igor or Vladimir or Ivan?). Accordingly, Shields trained animals in a paired-associate (PA) task in which links were established in memory between arbitrary sample-target pairs of nine-point polygon shapes. Using shapes dodged the limitation that animals do not know about politicians. On one screen of the task, animals saw the sample and the Star. If they chose the sample (a judgment of knowing), they were tested earning food rewards or long timeouts on the subsequent screen by having to choose between the associated target for the sample and a foil (another sample's target). If they chose the Star (a judgment of not knowing), they were tested on the subsequent screen with the possibility of reward but no risk of a long timeout. Testing the animals even after they responded Uncertain mirrored the important aspect of human metamemory experiments that humans attempt recognition both after feelings of knowing and after feelings of not knowing. Comparing these two performance levels lets one confirm that participants are uncertain when they respond Uncertain. This feature of this experiment was balanced against the risk that this would make the function of the Uncertain response more difficult for animals to grasp or its use too attractive. The idea in the experiment was that the sample would be the query that prompted the animal to judge whether it knew the target (i.e., the linked memory material). Therefore, a critical feature of Shields' task was that sample-target pairs occurred at different repetition rates so that some became better learned than others.

Table 1 (Row 6) shows a monkey's performance in this procedure. The sample-target pairs occurred at different repetition raters are treated as Difficulty Levels 1 to 4. The monkey learned more poorly the sample-target pairs that occurred more rarely. But he was unable to decline those trials selectively based on seeing the sample alone. Nor was there any difference in performance between the trials he chose to accept or to decline. Shields measured performance on chosen and declined trials separately, whereas Hampton measured performance on chosen and forced trials (the latter a combination of trials that would have been chosen and declined). To make Table 1 uniform throughout, we combined the chosen and declined performance levels algebraically into an estimate of the monkey's performance had he been forced to complete some memory tests. Thus we set Forced Percent Correct = Percent Chosen × Percent Correct Chosen + Percent Declined × Percent Correct Declined.

Thus the monkey showed no evidence of making Uncertain responses that were based in metamemory. Given this failure, Shields (1999) adjusted the procedure. She let the second test screen contain the sample at the top, the target and foil below it, and the Star below them. Now the monkey had visibly present all the information he needed to judge whether he should accept the memory test. Now he did (Row 7) respond Uncertain more to less-well-known sample-target pairs and he did show a performance advantage when he chose to accept the memory test. For some reason, though, the complexity of the PA task made it difficult for him to show this data pattern based on seeing the sample alone. It remains a research challenge to evaluate whether this is a real limit on animals' monitoring capacities, or whether something in Shields' procedure produced it. In any case, Shields' experiment suggests constructive ways to bring animal and human metamemory assessments closer together.

## 12. Tests of uncertainty monitoring in less cognitively sophisticated species

In turn, this result suggestive of limits on monkeys' metacognitive capacities makes one wonder about the monitoring capacities and limits of other species. Several researchers have looked for metacognitive capacities in species such as rats and pigeons that are less cognitively sophisticated and more associative in their behavioral solutions. Several preliminary comments are worthwhile here. These studies do not provide strong evidence for metacognitive capacities in these species. These negative data patterns could be important theoretically for suggesting that these species are not self-reflective in the way required to monitor uncertainty states and to respond adaptively to them. The phylogenetic map showing the emergence of self-reflective cognition in some species and not others would be exciting to see. However, one must also remember to interpret these null findings cautiously, considering whether the experimenters did enough of the right things for long enough to create conditions that ought to have let animals show a metacognitive capacity if they have it. This is a high hurdle for an experimenter to meet, because more extensive testing, or different motivators, and so forth, might have revealed the metacognitive capacity when the actual procedures did not. Our view is that these studies did a good job of arranging circumstances appropriately, so that these three studies provide negative results that may be interpretable and important. However, readers will wish to make their own judgment on this point.

### *12.1. Uncertain responses by rats in an auditory-discrimination task (Smith & Schull 1989)*

Smith and Schull (1989) adapted the psychophysical approach to studying uncertainty monitoring (Smith et al. 1995, 1997) by placing rats in a pitch-discrimination task. Animals were to make a left or right lever press, respectively, when they heard a repeating 400-Hz tone or a 400-Hz tone alternating with any other. A third response let the animals decline a trial at any time and begin a new trial instead. Animals knew well the effect of this response because they also used it to initiate trials. At first, participants were stabilized on an easy discrimination between 400-400 Hz tone pairs and 400-700 Hz tone pairs. Then the difficulty of the discrimination was increased by decreasing the frequency of the higher pitch until performance faltered at tone pairs of about 400-410 Hz.

Figure 8 summarizes the performance of six rats. On alternating trials, the alternating response predominated, whereas the repeating response predominated on repeating trials and on the most difficult alternating trials. Where these two response curves cross, the primary discrimination was performed at chance. The rats mastered well the two primary discrimination responses. But they did not recruit the Uncertain response adaptively at their threshold, as humans, monkeys, and a dolphin did. The rats did not assess when they were at risk for error in the primary discrimination and decline those trials selectively and adaptively.

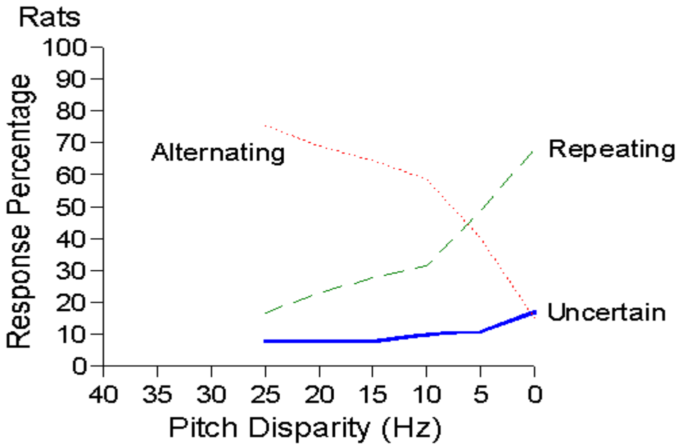


Figure 8. Performance by six rats in the frequency discrimination used by Smith and Schull (1989). The horizontal axis indicates the frequency difference between the alternating pitches on a trial. The repeating response was correct for trials with a frequency difference of 0, and these trials are represented by the rightmost data point for each curve. All other trials deserved the alternating response. The solid line represents the percentage of trials receiving the Uncertain response at each pitch disparity. The percentages of trials ending with the repeating response (dashed line) or alternating response (dotted line) are also shown.

We wondered whether the animals' problem was that the Uncertain response doubled as a trial-initiation response. To check on this, we incorporated an additional trial type (a higher tone repeating faster) that was rewarded randomly so that animals could only be 50% correct on this trial type (just as they are at threshold). The rats responded Uncertain to these trials three times as often as they responded Uncertain at threshold. This shows that they were associating some trials with lean reinforcement and that they were prepared to wave off these trials. In fact, it is interesting that animals declined the objective, stimulus-borne 50:50 contingency, but not the subjective, threshold-borne 50:50 contingency. This could suggest that the latter, subjective cue runs more quietly and less accessibly in the rat's cognitive system, whereas it seems lively and accessible in the cognitive systems of humans, monkeys, and dolphins.

We do not conclude from this that rats cannot monitor their uncertainty and decline trials based on it. Under different circumstances or methodology they might (e.g., if one combined a task in their strongest sensory modality with contingencies that strongly favored the use of the Uncertain response). However, Smith and Schull's version of the experiment continued for months while every day we tried unsuccessfully to coax the monitoring capacity from these animals. Perhaps it is difficult to document a sensitivity to these uncertainty or metacognitive cues because they are difficult for the rat to sense and use. This same difficulty arises regarding pigeons' cognitive systems, too.

### *12.2. Evaluating pigeons' capacity for metamemory in a DMTS task (Teller 1989)*

Teller (1989) was the first to study metamemory in the pigeon. As an undergraduate thesis, this project deserves special mention for its original contribution. Its methodology foreshadows strikingly that in Hampton (2001) and Inman and Shettleworth (1999 – to be described next). Six animals participated in a DMTS task with trials containing either 0-s or about 28-s forgetting intervals. There were two reinforcement schedules. By one discriminative response, the bird chose to complete the memory test with no hint information (i.e., without the sample highlighted on the screen as the obvious correct answer), receiving either no reward or a large grain reward. By another discriminative response, the bird chose to complete the memory test with the hint provided, almost surely receiving a small grain reward. On 60% of trials, birds chose between these schedules. When they chose the no-hint schedule, their performance is comparable to performance on Hampton's chosen trials. On 40% of trials, the animals were required to operate under either the hint or no-hint schedules. The forced-hint trial type balances the design but is uninteresting. The forced-no hint trial type is important. Performance on these trials is comparable to performance on Hampton's forced trials.

Under the hypothesis of metamemory, pigeons should choose the hint schedule (i.e., decline the memory test) more often after the delay. And they should perform better when they choose to complete the memory test than when they are forced to. The first prediction was confirmed (Table 1, Row 8 – the 0-s and 28-s delays are treated as Difficulty Levels 1 and 4, respectively). Remember, though, that this could reflect only that the choice of schedule has come under the associative control of the delay length. The birds were not more accurate when they chose the no-hint schedule than when they were forced to use it, as they should be if their choice was based on the monitoring of a more available memory trace. Their performance was similar to that of Hampton's Monkey 2 (Row 2), who also showed a minimally metacognitive data pattern. Teller concluded that his results did not demonstrate a capacity for memory monitoring in pigeons, though he suggested that future research might be able to do so and focused his discussion on methodological considerations regarding future research.

### *12.3. Evaluating pigeons' capacity for metamemory in a DMTS task (Inman & Shettleworth 1999)*

Inman and Shettleworth evaluated pigeons' metamemory capacity using a similar but independent approach. In Experiment 1, four birds saw one of three possible samples followed by a forgetting interval of 1-8 s. On 33% of trials, a normal DMTS memory test (that could earn a large food reward) followed the delay. These trials, on which the animal was forced to complete the memory test, correspond to the forced trials in Teller and Hampton. On 33% of trials, the delay ended with a safe response that earned a small food reward. As in Teller's case, these trials balance the design and are uninteresting. On 33% of trials, the delay ended with the safe response and the DMTS memory test presented in combination, so the animal could choose whether to decline or accept the test. The trials accepted of these combined trials correspond to the chosen trials in Teller and Hampton. The familiar predictions from this experiment (under the metamemory hypothesis) are 1) the use of the safe key to decline memory tests should increase at longer delays; and 2) there should be a performance advantage for chosen over forced memory tests, especially at longer delays, as one of Hampton's monkeys showed but as Teller's pigeons did not show.

Inman and Shettleworth observed (Table 1, Row 9) that longer delays only produced a 7% increase in the use of the safe response (not significant by a parametric test), even though the birds performed worse at the longer delays and should have used that response more (as Hampton's and Teller's animals did). Moreover, there was no reliable performance advantage for chosen trials over forced trials, as one of Hampton's monkeys clearly showed, and there was no interaction between the length of the delay and the size of this advantage, as should be true under a metacognitive hypothesis.

In Experiment 2, Inman and Shettleworth asked four pigeons to choose between accepting the memory test and declining it (via the safe response) after the forgetting interval but before the choice objects were revealed. This experiment is almost identical to Hampton's monkey experiment, but with different results (Table 1, Row 10). Again longer delays only produced a small, nonsignificant increase (5-7%) in the use of the safe response. Again there was no advantage for chosen trials over forced trials, and there was no interaction between the length of the delay and the size of this advantage.

Inman and Shettleworth concluded appropriately that their data did not show that pigeons used memory-trace strength as a discriminative stimulus, though they acknowledged it remained possible birds could do so (e.g., if a wider range of matching accuracies was sampled, perhaps by increasing forgetting intervals beyond 8 s). Like Smith and Schull with rats and Teller with pigeons, Inman and Shettleworth suggested that pigeons might have only a weak metamemory capacity, making it difficult for the experimenter to observe it because it is difficult for the birds to use.

## 13. A unifying formal perspective

Now we offer a unifying formal description of performance in the tasks reviewed here. This description serves several useful ends. It supports cross-task and cross-species comparisons among data patterns. It allows the studies of optimality that figure prominently in discussions of animal behavior. It offers a neutral description of performance that is inclusive theoretically because it makes no theoretical commitments toward behaviorism or cognitivism. It clarifies the formal structure of behavior so that different theoretical perspectives can be brought to bear on it.

Our model is grounded in Signal Detection Theory (SDT – MacMillan & Creelman 1991). SDT assumes that performance in perceptual or memory tasks is organized along an ordered series (a continuum) of psychological representations of changing impact or increasing strength. In the same-different task, for example, the continuum of subjective impressions would run from clearly different (a large disparity between stimuli) to same (zero disparity – the X-axis in Figure 9a). Given this continuum, SDT assumes that an objective event will create subjective impressions from time to time that vary around some average impression. A threshold different trial might create impressions that vary as Figure 9a's D (Different) normal distribution. Same trials might create impressions that vary as Figure 9a's S (Same) normal distribution. The overlap between these distributions is what ensures errors and fosters uncertainty in the task – both kinds of trials can feel alike to the perceiver. Finally, SDT assumes a decisional process by which criterion lines are placed along the continuum so that response regions are organized. In Figure 9a, as stimulus pairs made disparity impressions that fell to the left of the Different-Uncertain criterion line, to the right of the Uncertain-Same criterion line, or between these two, the participant would make the Different (D), Same (S), or Uncertain (U) response, respectively.

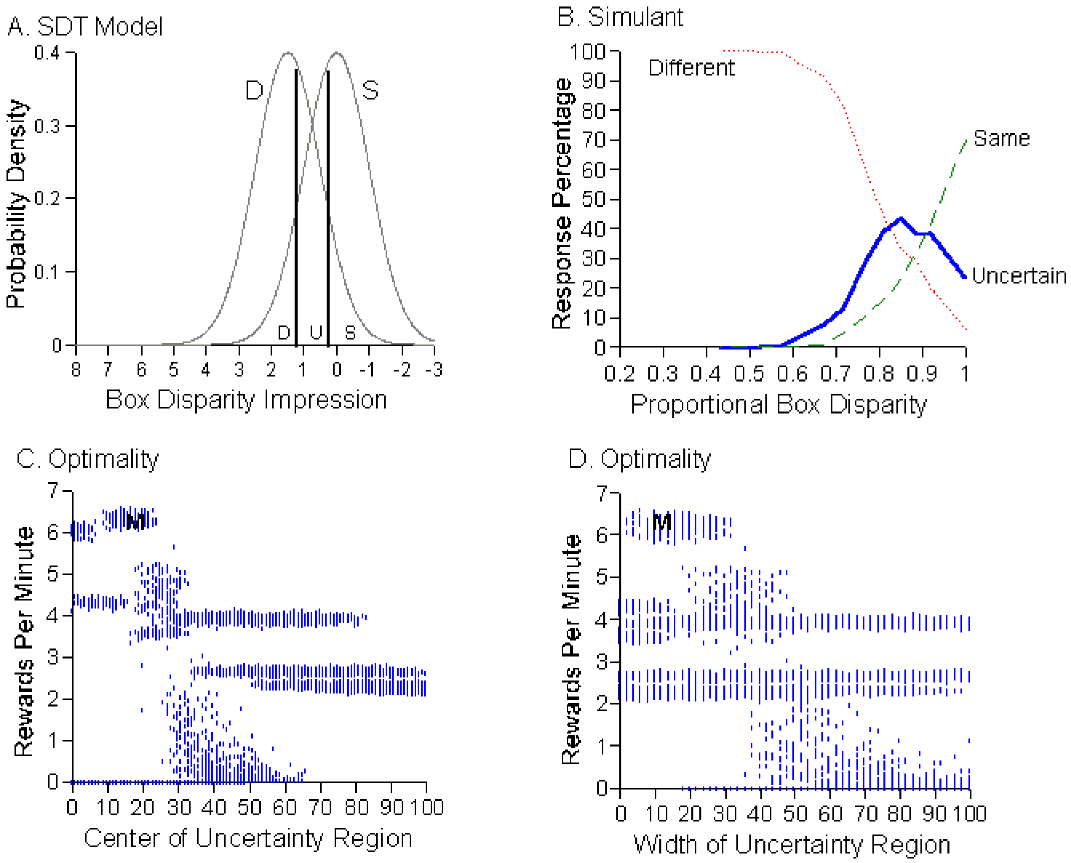


Figure 9. A. A signal detection theory (SDT) portrayal of monkeys' decisional strategy in the same-different task of Shields et al. (1997). Unit-normal disparity-impression distributions are centered at 0.0 for Same (S) trials and at a positive disparity for Different (D) trials. These normal curves are overlain by decision criteria that define the animal's three response regions (from left to right, Different [D], Uncertain [U], and Same [S]). B. Performance by the simulant that fit best the monkeys' performance (compare Figure 6b) in the same-different discrimination task of Shields et al. (1997). The horizontal axis gives the ratio between the densities of the comparison box and the standard box for trials of different disparities. The same response was correct for trials at a proportional box disparity of 1.0 – these trials are represented by the rightmost data point for each curve. All other trials deserved the different response. The solid line represents the

percentage of trials receiving the Uncertain response at each density ratio. The percentages of trials ending with the same response (dashed line) or different response (dotted line) are also shown. C. The reward efficiency (in rewards earned per minute) of simulants that centered the Uncertain response region at different places along the disparity-impression continuum in a virtual version of Shields et al.'s (1997) same-different discrimination. We surveyed the reward efficiency of 5,151 decisional strategies when each received 8,000 trials in a simulated version of the task, subject to the trial times, penalty times, and reward structure of the task the monkeys experienced, and using the signal-detection response rule that accorded with the three response regions defined by each simulant's two criterion placements. M represents the position in this optimality space of the simulant that best fit the performance of the real monkeys. D. The results of the same simulation plotted by the width of the Uncertain response region.

We will illustrate the application of the SDT model to animals' performances in the same-different (SD) task of Shields et al. (1997), the serial probe recognition (SPR) task of Smith et al. (1998), and the delayed matching-to-sample (DMTS) task of Hampton (2001). The details of the simulations are given in Appendix 1.

### *13.1. The SD task of Shields et al. (1997)*

The goal of applying the SDT model was to find the performance parameters of the simulated perceiver who – while conforming to the model – produced performance most like that which the monkeys showed (Figure 6b). The crucial step in modeling was to assess the decisional strategy that monkeys probably used by sampling many different placements for the Different-Uncertain and Uncertain-Same criteria shown in Figure 9a. In fact, we modeled the performance of 520,251 simulated perceivers (hereafter simulants) who had different decision criteria. Figure 9b shows the performance of the simulant among these whose performance most closely matched that of the monkeys (compare Figure 6b). The predicted response percentages were within 3-4% of their observed targets. This fit compares favorably with the fit of other formal models in the experimental literature (e.g., Smith & Minda 1998, 2000). The best-fitting simulant placed its criteria at .825 (Different-Uncertain) and .905 (Uncertain-Same) along the subjective-disparity continuum.

The SDT model also lets us assess the optimality of the monkeys' decisional strategy. To do so, and to illustrate methods in this area, we evaluated the reward efficiency of 5,151 simulants who performed the SD task using variously placed decision criteria while also experiencing virtually the trial times and punishment times animals experienced. That is, we calculated the rewards per minute that each decisional strategy would earn. Figures 9c,d show the rewards per minute earned by simulants that centered their Uncertain response region at different places along the disparity continuum and that gave this region different widths. M denotes the monkeys' best-fitting simulant. The monkeys' decisional strategies were essentially optimal because they centered their Uncertain response region at their threshold for discriminating same from different and because they widened it judiciously, too.

We point out that this optimality study, like the SDT model it is based on, is psychologically neutral regarding the processes and representations that underlie performance and regarding the level in the cognitive system these processes and representations occupy. Animals could regulate optimally using low-level, high-level, or even conscious processes. Animals could regulate poorly at these levels, too. It is a step beyond finding animals' positions in an optimality landscape to judge the cognitive sophistication of the decisional strategy that placed them there.

### *13.2. The serial-probe-recognition (SPR) task of Smith et al. (1998)*

In this case, SDT would assume that the list items create subjective memory impressions that lie along a continuum of trace strength (the X-axis in Figure 10a). The probe then queries the strength of one trace. Probes on Not There trials will generally point to weak traces, perhaps averaging 0.0 plus or minus the scatter of memory variability (the normal distribution NT in Figure 10a). Probes on There trials will point to stronger traces on average (though still with memory variability), especially for the primacy and recency list items (the four T normal distributions in the figure). The overlap between the Not There and There distributions is what makes the SPR task difficult and uncertain.

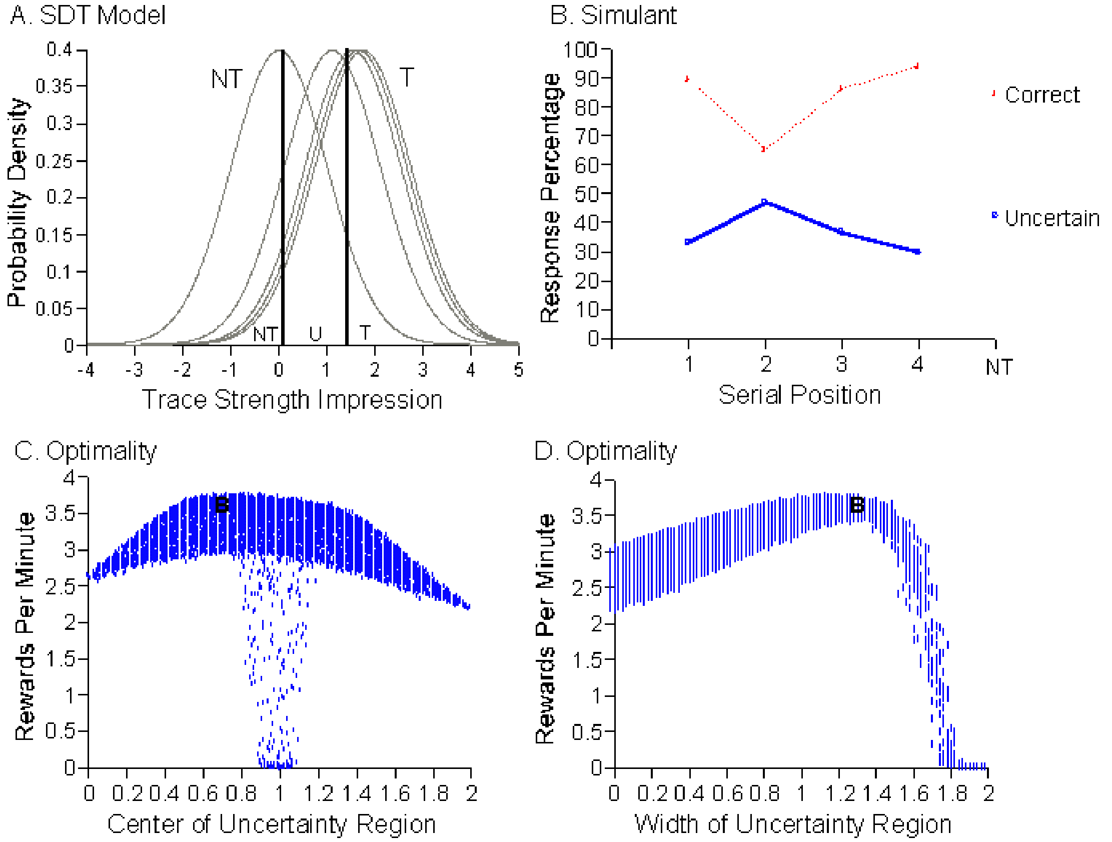


Figure 10. A. A signal detection theory (SDT) portrayal of Monkey Baker's decisional strategy in the serial probe recognition task of Smith et al. (1998). Unit-normal trace-impression distributions

are centered at the locations along the trace-strength continuum corresponding to the animal's d' for probes of the four serial positions in the memory lists (T), and at 0.0 for the not there probes (NT). These normal curves are overlain by the decision criteria that define the animal's three response regions (from left to right, Not There [NT], Uncertain [U], and There [T]). B. Performance by the simulant that fit best Monkey Baker's performance (compare Figure 7a) in the serial probe recognition task of Smith et al. (1998). NT denotes Not There trials. The serial position (1-4) of the probe in the list of pictures is also given along the X-axis for the probes on There trials. The percentage of total trials receiving the Uncertain response is shown (solid line). The percentage correct (of the trials on which the memory test was accepted) is also shown (dotted line). C. The reward efficiency (in rewards earned per minute) of simulants that centered the Uncertain response region at different places along the trace-strength continuum in a virtual version of Smith et al.'s (1998) serial probe recognition task. We surveyed the reward efficiency of 5,151 decisional strategies when each received 8,000 trials in a simulated version of the task, subject to the trial times, penalty times, and reward structure of the task the monkeys experienced, and using the signal-detection response rule that accorded with the three response regions defined by each simulant's two criterion placements. B represents the position in this optimality space of the simulant that best fit the performance of the real Monkey Baker. D. The results of the same simulation plotted by the width of the Uncertain response region.

Once again SDT assumes a decision process by which criterion lines are placed along the continuum to define response regions. In Figure 10a, as a probe stimulus contacted a trace that fell to the left of the Not There-Uncertain criterion line, to the right of the Uncertain-There criterion line, or between these two, the participant would make the Not There, There, or Uncertain response, respectively.

To find the best-fitting configuration of the SDT model, we modeled the performance of 226,981 simulants with differently placed decision criteria. Figure 10b shows that the performance of the best-fitting simulant closely reproduced Baker's performance (compare Figure 7a). On average, the response percentages were within 3% of their observed targets. Figure 10a shows the SDT description of this simulant.

We also drew the optimality landscape of Smith et al.'s SPR task. To do so, we evaluated the reward efficiency of 5,151 simulants that centered the Uncertain response region at various places along the trace-strength continuum and that widened it to varying degrees. Figures 10c,d show the rewards per minute of simulants that gave the Uncertain response region different centers and widths. B indicates the position in these landscapes of Monkey Baker's best-fitting simulant. Here, too, the monkeys centered and widened their Uncertain response region adeptly. As Figure 10a shows, they declined those trace strengths that were most indeterminate and that most risked error.

Optimality studies like these can support one's experimental planning in this area. They let one preview how different penalties and rewards change the shape of the optimality surface, and they let one find experimental settings that emphasize the value of the metacognitive strategy over alternatives. This may encourage animal participants to adopt the metacognitive strategy if they can. This preview may be especially important in the domain of comparative metacognition for this reason. We believe that the metacognitive strategy is a subtle and effortful one even for monkeys. In our experience, even monkeys gravitate toward an associative, nonmetacognitive performance strategy if they can find an effortless one that earns a decent rate of return. Hampton's Monkey 2 may illustrate this tendency. An experiment that creates the maximal separation between the metacognitive and nonmetacognitive strategies on the optimality surface may help convince monkeys that the cognitive effort of the former is worthwhile. For rats and pigeons, this maximal separation favoring metacognitive monitoring may be even more critical because these species have difficulty expressing the metacognitive capacity at all.

### *13.3. The DMTS task of Hampton (2001)*

Regarding Hampton's Monkey 1, we assumed (Figure 11) that samples left behind strong or weak impressions after short or long delays (the four S [Sample] normal distributions), and that foils (non-sample items) were represented by weaker memory impressions that averaged 0 with memory variability (the F [Foil] normal distribution). We also assumed that the animal placed one criterion line on this trace-strength continuum to separate the Decline and Accept response regions. In this way he would decline or accept memory tests as the monitored trace was weaker or stronger than this Decline-Accept criterion. Notice how the crucial predictions from memory monitoring emerge given this model. First, the lower trace strengths given longer delays will mean more below-criterion sample traces and thus higher percentages of trials declined. Second, if the animal evaluates the availability of the sample's trace in memory, then he can selectively choose to accept memory tests when he monitors strong sample traces and he will perform especially well on these trials he accepts.

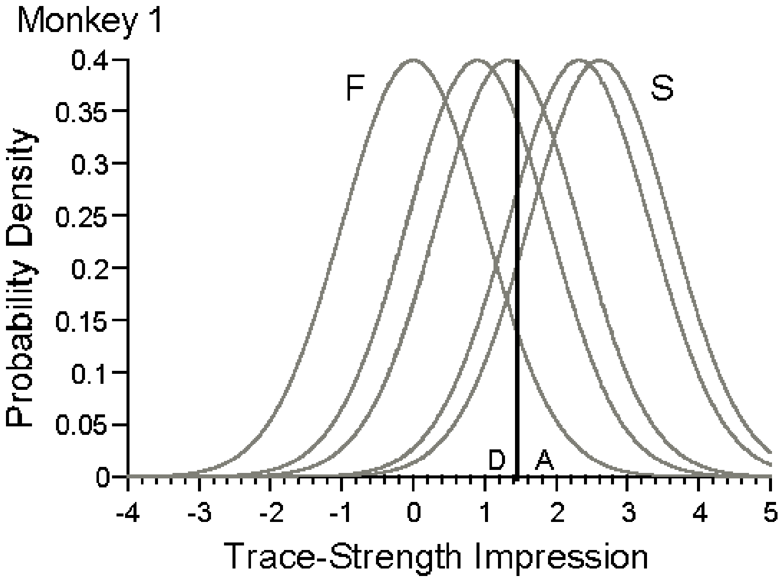


Figure 11. An SDT portrayal of Monkey 1's decisional strategy in the delayed matching-to-sample task of Hampton (2001). Unit-normal trace-impression distributions are centered at the

locations along the trace-strength continuum corresponding to the animal's d' for samples remembered after four forgetting intervals (S), and at 0.0 for foils not seen on that trial (F). These normal curves are overlain by the decision criterion that defines the animal's two response regions (from left to right, Decline [D] and Accept [A]).

To find the best-fitting parameter configuration of the SDT model, we modeled the performance of 201 simulants that had differently placed Decline-Accept criteria. Table 1 (Row 3) shows that the best-fitting simulant reproduced well the observed performance (Row 1). The predicted response percentages were within 3-4% of their observed targets. Figure 11 shows an SDT description of this simulant.

It is worth comparing (Figures 10a , 11) the SDT descriptions of the memory-monitoring performances achieved by Baker and by Monkey 1. Monkey 1, of course, had no Not There response region and so lacks the second, lower criterion point. But the decision he had to make (whether the monitored trace was strong enough to accept the memory test) is like the second decision that Monkey Baker had to make (whether the monitored trace was strong enough to respond There). It is interesting that across laboratories, methodologies, and monkeys this upper criterion lies at the same place on the trace-strength continuum. Further work could pursue similarities like these and possibly establish the point of confidence at which animals think they know or remember. Meanwhile, this similarity suggests that Baker and Monkey 1 were responding in a similar way to the same memory-strength cue.

### *13.4. Formal modeling and theory development*

Our hope is that formal models like these could eventually support theory development in this area. We illustrate the kind of theoretical insight they might offer. There are two memory-monitoring strategies that Monkey 1 could have used in Hampton's experiment. By one strategy, based in short-term memory, the animal would "rehearse" the sample as best he could, and after the delay assess whether that one trace was still available enough to accept the memory test. By another strategy, based in longer-term memory, the animal would passively wait out the delay and then query all four relevant memory locations to see whether any trace was available enough to accept the memory test (for Hampton's monkeys, only four stimuli were relevant at a time).

Which of these strategies was Monkey 1 using? Table 1 (Row 3) shows the simulant that best-fit Monkey 1's performance while using the short-term strategy of rehearsing and monitoring only the sample's trace. Performance on chosen trials hardly decreases as the delay gets longer. This happens because the animal only accepts the test if the sample's trace is available enough, and so the sample's trace when he accepts is always about the same amount above the foil distribution, keeping him equally far from error.

Table 1 (Row 4) shows the simulant that best-fit Monkey 1's performance while using the long-term strategy of monitoring activity at the four relevant memory locations. Performance on chosen trials drops sharply at longer delays. This happens because this simulant accepts the memory test if any trace is available enough. With longer delays, the active-enough trace is more often a falsely-active foil, and on these occasions the simulant will choose to complete the memory test but will err. One sees from Rows 3 and 4 that formal models can differentiate task strategies and thus can be used to suggest additional research and advance theory.

Monkey 1's performance (Row 1) seems to combine these two patterns. Out to Difficulty Level 3 (a 50-s forgetting interval), his performance on chosen trials stays high, recalling the short-term strategy shown in Row 3. At 100 s, his performance on chosen trials has fallen sharply, recalling the longer-term strategy shown in Row 4. It is possible that this interpretation – of a transition between memory strategies at long delays – could be supported using what is known about the temporal limits of monkeys' working memory (Fobes & King 1982). We offer it only to show that models may illuminate subtle but theoretically important processing differences in animals' performances in uncertainty-monitoring tasks. Actively asking whether the target trace is still available enough, and simply waiting to query the small set of relevant memory traces, are interesting but different capacities.

### *13.5. Formal modeling and alternative theoretical perspectives*

We chose the detection framework as a constructive way to understand animals' performances. However, by assuming trace strengths that quantitatively fade with time, or that wax and wane with serial position, we hope we do not foreclose interest in other theoretical approaches. For example, psychologists have debated whether human recognition memory or memory retrieval can be explained using quantitative continua like trace strength, or whether one must also assume the contextually bound states that are typical of conscious episodic memories and that are often qualitatively present or absent. One could model Hampton's data from Monkey 1 using a more episodic approach, by assuming that at the end of the forgetting interval the animal queried whether it still had qualitatively available the memory of seeing the sample. This approach would also explain in an interesting way why this animal never accepted any trials when, in another condition, Hampton gave him memory tests without having shown him any sample. We point out that Hampton was not willing to endorse this strong a metacognitive interpretation of the performance of Monkey 1. We also point out that this interpretation works less well for the SPR data reported by Smith et al. Their animals might have responded There when they found a contextually bound memory for the previous presentation of the probe item, and might have responded Not There when they did not. However, one still must explain the animals' Uncertain responses which suggest that some quantitative assessment of memory was at work, too. One could speculate that the animals were uncertain whether they had recovered a contextually bound memory, but, as in Hampton's case, the data do not require this interpretation.

In this emerging field, different formal perspectives may each make constructive contributions. Our quantitative approach raises one set of questions about animals' decision-making and criterion-setting processes and about the level these processes have in the psychological system. The qualitative approach would raise another set of questions about animals' episodic memories. In fact, the problem of showing episodic and possibly conscious recollection in animals is an active one in comparative psychology (Menzel 1999; Schwartz et al., 2002).

## 14. The psychology of uncertain responses and uncertainty-monitoring performances

The facts regarding animals' performances are clear and the SDT descriptions neutral because they do not assume processes (high- or low-level) that would suit one branch of comparative theory but not another. Now, though, we come to the difficult point of considering the appropriate psychological description of performance. What is the psychological character of the representations that underlie the SDT continua? How do animals place decision criteria along a continuum and use them? These questions go into areas where different theoretical perspectives are preferred and where theoretical tensions may spark. Yet finding the correct psychological description is one crux of the matter. Therefore, in Sections 14.1-14.4, we offer some considerations that help us think psychologically about animals' performances and about their use of Uncertain and Decline responses.

### *14.1. Animals' Uncertain and Decline responses are not associatively based and are not under stimulus control*

One consideration is that a variety of stimulus-based, low-level interpretations of animals' performances are untenable. For example, in the memory-monitoring tasks of Smith et al. (1998), all stimuli, across trials, became targets and foils and were rewarded and non-rewarded following both primary responses. No stimulus cue indicated any response. Only presence or absence in the memory list was relevant. The psychological action in this experiment likely occurred along an internal continuum of subjective trace strength or availability, with animals declining memory tests when probes encountered memories of indeterminate strength. These trial-by-trial assessments of memory strength are profoundly different from the signals available in traditional operant situations, leaving the monkeys' behavior in this task far from traditional senses of stimulus control.

This is already an important interpretation and constraint on psychological theorizing. In essence, the animals are monitoring the contents of memory on each trial. In essence, they are being metacognitive, though one may not wish to load the animals' performances with all the theoretical baggage that this term has with humans.

A similar conclusion applies to Shields et al.'s (1997) SD task. Behavior in this task also cannot be controlled by absolute stimuli, whether through the generalization gradients surrounding them or the reinforcement histories associated with them. Indeed, the abstractness of the same-different judgment is why many species cannot make it. Here, too, the cues the animals used to decline trials must have been cognitively derived, representing a decision about the (indeterminate) status of the relation assessed on each trial.

### *14.2. Metacognitive performances and the parsimony embodied in Morgan's Canon*

But does this conclusion extend to the perceptual-threshold tasks that began the exploration of animals' metacognitive capacities (Sections 6 and 7, Smith et al. 1995, 1997)? Here one could explain performance using stimulus control or reinforcement history. For example, one might say that stimuli of intermediate density were mildly aversive for being associated with errors and timeouts and that Uncertain responses were conditioned in these stimulus contexts. This low-level explanation has a distinguished pedigree. It defends the principle of parsimony embodied in Morgan's Canon (1906. p. 53) that grants animals only simple cognitive capacities. (Remember that Morgan's idea was that one should always interpret an organism's behavior at the lowest possible psychological level.) Thus, given metacognitive-like performances by animals, there is a 100-year-old urge to knock them down, to dismiss them as low-level associative phenomena. Readers may have felt this urge as they read this article. We have several cautions about this theoretical position.

For one thing, in this case the parsimony embodied in Morgan's Canon may be false. The problem is that one cannot interpret the animal's "metacognitive" performance in a vacuum. Humans perform the same way – indeed, the graphs shown in Figure 3 and Figure 4 present some of the strongest existing parallels between human and animal performance. Humans report that they are consciously uncertain and reflexively self-aware as they produce these graphs. And, humans and monkeys share evolutionary pasts, adaptive pressures, and homologous brain structures. Thus it is unparsimonious to interpret the same graph produced by humans and monkeys in qualitatively different ways – consciously metacognitive vs. low-level associative. It uses two opposed behavioral systems to produce the same phenomenon when one might do. Moreover, this duality of interpretation may even be an inappropriate scientific stance. In any other domain, if you showed identical graphs, and then nonetheless offered qualitatively different high-level and low-level interpretations, you would be howled out of the journal. If it were older and younger children, or young and aged adults, or individuals without and with depression, you would have no warrant to do so. Likewise, in the case of humans and animals, you may have no warrant to do so, either.

Given identical performances by humans and animals, one could take the consistent but radical step of claiming that animals and humans are controlled by stimuli and reinforcement in these perceptual tasks, and that humans' introspections and reports of reflexive consciousness and metacognition are a non-functional epiphenomenon. This step, though, would deny 100 years of introspections by humans in these tasks, deny the careful understanding that early psychophysicists reached about the Uncertain response in these tasks, and deny the primary source of evidence (self-reports and introspections) that we have for human metacognition and even consciousness. This is a lot to pay for reserving the right to dismiss animal minds.

Fortunately, the issue need not be to either elevate animals or denigrate humans. Rather, given true comparative data and identical performances by humans and animals, our point is just that a reasoned, middle descriptive ground is preferable to the clash between divergent explanatory frameworks at different psychological levels. It is an important principle of comparative research that the kind of integrative parsimony and simplicity of explanation one seeks when explaining the performance of several species in several tasks will be different from the parsimony one seeks when explaining the performance of a single species in a single task. The single-species character of much of comparative psychology has encouraged a sharp focus on low-level, associative kinds of parsimony – the parsimony of Morgan's Canon. Multiple-species studies might foster interest in more integrative kinds of parsimony that could make complementary contributions to theory in the comparative literature. In the present case, a reasoned middle ground could consider both common processing principles across species and acknowledge possible experiential differences across species in whatever way the whole empirical picture warranted.

For another thing, remember that the perceptual-threshold results do not exist alone. Once one knows that animals are using the Uncertain response adaptively to decline derived cognitive states (e.g., indeterminate Same-Different relations, indeterminately available memory traces), another issue of parsimony arises. For now if one explains animals' perceptual performances using a low-level mechanism, but one must explain their memory or same-different performances as a more sophisticated kind of cognitive monitoring, then one grants them two different indeterminacy-resolution systems, one of which is already fairly high-level. In this case one could explain the data more simply just by invoking one indeterminacy-resolution system that applies to indeterminate memory traces, threshold perceptual impressions, ambiguous relational judgments, and possibly many real-world situations, too. Why is such an indeterminacy-resolution system so implausible a thing to assume that animal minds might have benefited from having (Section 15)? Why is such a system assumed to be more complicated psychologically than if animals learned many sequential reinforcement histories along a perceptual continuum? The history of the comparative literature has led it to answer these questions in one way when analyses of the psychological structure of the different capacities might answer these questions in another way.

Finally, remember also our failure to teach rats to decline trials near their auditory threshold. Rats, pigeons, and other associative species could certainly learn a reinforcement contingency attached to a middle stimulus class between two others and could have a response brought under the control of that class of stimuli. If this were all there is to the threshold task, rats would escape from threshold trials naturally and easily. They do not. This also rules out that these tasks are simply about middle-stimulus avoidance. The threshold tasks seem to be psychologically structured in some way that leaves rats out (insofar as the methods were sufficient to elicit the crucial capacities from them) but leaves humans, monkeys and dolphins in. Apparently these latter three species are using a capacity at some higher level in the cognitive system that rats access with difficulty.

### *14.3. Tests of uncertainty monitoring provide inconsistent mappings between stimulus inputs and behavioral outputs, and therefore encourage controlled cognitive processes*

Fortunately, one can analyze the psychological structure of the tasks considered in this article, and advance toward a descriptive middle ground that explains indeterminacy-resolution systems across tasks and across species. Shiffrin and Schneider (1977, pp. 167-168) considered the information-processing consequences of the ambiguity that arises from indeterminate cognitive processing. In their description, indeterminate mental representations of stimuli necessarily mapped inconsistently and unreliably onto behavioral responses, making those representations poor indicators of what the organism should do. As a result, higher levels of controlled processing were invoked to adjudicate the indeterminacy. All the uncertainty monitoring tasks are inconsistently mapped and benefit from controlled processing in this sense.

For example, the perceptual-threshold tasks deliberately challenge the observer's discrimination ability. Given perceptual error, true dense (e.g.) trials and threshold sparse trials will often produce the same subjective impression. Consequently, the impression of density itself will not indicate reliably the appropriate response, and higher-level mediation will be valuable to compare the impression to existing decision criteria to choose a behavioral option. From this perspective, the Uncertain response would be one manifestation of near-threshold resolution processes. It might represent a decision that the trial should be declined because the primary discriminatory process had failed. Note that this description applies to both human and animal cognition. Psychophysical procedures ensure indeterminate stimulus-response mappings and encourage controlled decision-making processes no matter the participant species. This realization provides a theoretically gentle way to grant animals' Uncertain responses some of the cognitive sophistication that is due them.

As another example, given memory error in the SPR task, the same trace strength or availability monitored on a trial could be caused by either a there probe or a recently seen not there probe, and so the trace strength monitored cannot reliably tell the animal what to do. Once again these inconsistent representation-response mappings create the need for controlled resolution processes (see also Gilden & Wilson 1995). In fact, regarding recognition memory specifically, Atkinson & Juola (1974) suggested that the range of indeterminate trace strengths may require qualitatively different, secondary information-processing strategies (i.e., an extended memory search). These indeterminate trace strengths are just the ones that monkeys and humans decline selectively. Their Uncertain responses probably represent a controlled decision, on the threshold of recognition, not to complete the memory test.

To clarify the distinction between decisionally controlled processes and stimulus-controlled processes, consider the conditional discrimination that many humans perform daily – green-go; red-no go. These distinct stimulus input classes eliminate mistakes in perception – stimulus impressions map consistently and reliably onto appropriate responses. Fine discriminations, decision making, and decision criteria are irrelevant. Stimulus and response may associate so strongly that responses are triggered automatically, reflexively, stereotypically, fast, and effortlessly. In fact, this is the point of the worldwide consistent mapping that stoplights offer.

In contrast, imagine that traffic lights gradually morphed from red to green, and that drivers decided whether their light was green enough to go. This situation would be about decision making and decision criteria. It would be about controlled cognitive processing that would be attentional, capacity intensive, and slow. It would also be a nightmare, as perceptual error and self-serving criteria made indeterminate lights seem green enough to go for orthogonal travelers. This kind of decisional task is the one that humans and animals face in the paradigms described in this article. They must decide whether the box is dense or sparse enough to try, the pitch high or low enough, the stimulus relation same or different enough, the memory trace familiar or unavailable enough. Even associative theorists, working in the behaviorist tradition, have given threshold situations like these special attention, for they find that the rules of stimulus control can change there, that animals become minimally informed observers there, and that animals have special problems finding adaptive solutions there (Boneau & Cole 1967; Commons et al. 1991; Davison et al. 1985; Miller et al. 1980; Terman & Terman 1972). In granting this special treatment they follow the classical psychophysicists, who saw the threshold state, and the Uncertain response, as distinctive and complex psychologically.

### *14.4. Animals may share with humans a theoretically important construal of the threshold and memory-monitoring tasks*

Humans make a particular construal of these tasks (i.e., the Dense-Sparse task, the Same-Different task, the serial-probe recognition task) that explains their performance and verbalizations. They accept that the tasks have two primary input classes (dense-sparse, same-different, there-not there). They accept that every trial presents one of these input classes and has a correct answer if they could just discern it. Thus, humans map the two stimulus input classes to the two primary responses, use these responses when they think they know, and reserve the Uncertain response for situations of indeterminacy.

Given this mapping, the Uncertain response alone has no input class associated with it. It is about the status of the primary discriminatory process and about its probable failure. It stands structurally outside the primary discrimination and intrinsically meta to it. For humans, it even attaches to declaratively conscious uncertainty states and moreover uncertainty states that are reflexively self-aware in the sense that humans say "I am uncertain." This task construal explains why the Uncertain response feels less stimulus-based, why it can feel like cheating or like mental shirking. It explains why the Uncertain response alone can be omitted from the task's grammar, and why some humans do so by an act of will or bravado. The other two responses cannot be omitted and no human would ever do so. This task construal also confirms the special psychological status that psychophysicists always granted the Uncertain response.

Now, on turning to the animals' performances, leave aside verbalizations, bravado, and consciousness. Animals could still share with humans their abstract construal of the perceptual and memory tasks. In fact, we believe animals do construe these tasks as having two primary input classes and two primary responses that directly map to one another and that exhaust the trial environment, leaving the Uncertain response with no input class associated with it and with a distinctive role in the grammar of the task. The animals make this task construal because we train them to, remind them to daily, and in some conditions force them to do so. First, we always train animals extensively in the primary discrimination before giving them an Uncertain response in the task. That response arrives as an optional, extra response, with the two primary input classes and the two primary response classes already established, mapped to one another, and sufficient for performance in the task. Second, we generally provide animals a warmup every day during which they receive easy discrimination trials that ramp up in difficulty toward threshold. During this warmup, it is clear that there are only two stimulus input classes and two useful kinds of responses, and animals only make these two responses. Third, we sometimes run animals in conditions in which they must perform the mature discrimination task without the Uncertain response available. This reinforces again that the task's two input classes and two responses are sufficient for performance in the task. Fourth, animals are also encouraged toward their construal of the task, and toward granting the Uncertain response a special role in the task, by the fact that the Uncertain response alone never earns a direct reward, never earns a timeout penalty, and always has the same neutral function and result in every trial context. Thus we believe our animals are massively trained in making just the construal of these tasks that we have been describing.

In summary, then, these considerations lay the groundwork for a psychological understanding of humans', monkeys', and a dolphin's successful performances in uncertainty-monitoring tasks. They offer a balanced psychological interpretation of Uncertain responses, granting them the cognitive sophistication they deserve without burdening them with consciousness or with equally heavy behaviorist assumptions.

## 15. Declarative consciousness

Notice, though, that these considerations do not imply that animals feel uncertainty in these tasks or evaluate within explicit consciousness the status of perception or memory. Regarding the role of declarative consciousness in these tasks, authors (especially ourselves) have been notably cagey. For example, Smith et al. (1998, p. 245) suggested that "one could scale back the claims of consciousness while preserving something of the sophisticated, memory-based, flexible, controlled mediational processes that do represent a higher level of choice and decision making and that are needed to explain how monkeys cope with (and escape) indeterminate memory events." Hampton's (2001, p. 5361) assessment of his analogous memory findings was similarly cautious. On the one hand, he concluded that "the ability of these monkeys to appropriately decline memory tests when they were unlikely to choose the correct image at test indicates that they know when they remember, a capacity associated with conscious cognition in humans." However, Hampton also pointed out (p. 5359) that it is "probably impossible to document subjective, conscious properties of memory in nonverbal animals." Thus the tack he took (p. 5359) was to stress that monkeys have "an important functional parallel with human conscious memory," or an "important functional property of human conscious memory."

This issue is actually more complex and interesting than it would be if one just prudently denied consciousness to animals as they perform the tasks described here. In fact, our research and that of Hampton raises the cherished idea in cognitive science that cognitive indeterminacy and difficulty inherently elicit higher-level and even conscious modes of cognition and decision making in the organism (Dewey 1934/1980; Gray 1995; James 1890/1952; Karoly 1993; Nelson 1996). James (1890/1952, p. 93) noted that consciousness provides extraneous help to cognition when nerve processes are hesitant. "In rapid, automatic, habitual action it sinks to a minimum." Dewey (1934/1980, p. 59) also argued that in habitual, well-learned behaviors the behavioral impulses are "too smooth and well-oiled to admit of consciousness." Tolman (1932/1967, p. 217) noted that "conscious awareness and ideation tend to arise primarily at moments of conflicting sign-gestalts, conflicting practical differentiations and predictions," such as when the animal is poised on the threshold of a difficult discrimination. We have already encountered Tolman's interesting claim that animals' uncertainty behaviors could operationalize consciousness. Karoly (1993, p. 25) emphasized that uncertain, conflicted conditions are the ones that initiate self-regulation. Gray (1995) described the special neural circuits that may arrest behavior, increase arousal and redirect attention and mental resources toward the causes of difficulty (see also Smith 1995). In the psychophysical literature on humans, too, there is a common view that criterion-setting mechanisms are consciously meta to the primary discriminatory process (Swets et al. 1961; Treisman & Faulkner 1984). Indeed, some psychophysical methods (e.g., the construction of receiver operating characteristic [ROC] curves) depend on humans' ability to obey explicit instructions and consciously set confidence criteria at appropriate levels. Humans comply with all this and often report that their setting of criteria along a continuous dimension is a strategic cognitive process aided by conscious self-regulation.

Now one could still interpret the human and animal results dualistically in this respect, by granting humans declarative, subjective consciousness in these tasks but animals only the unconscious, functional parallels to human conscious cognition. However, we suggest that the claim that difficulty and uncertainty elicit conscious modes of self-regulation is a plausible, principled claim about cognitive architectures generally, one that applies to the human species and to some animal species as well. In fact, this claim has behind it a strong adaptive pressure that might have led as follows to the evolution of working consciousness.

Normally, the systems of human cities (water, food, travel, safety) operate autonomously, reflexively, and automatically in highly trained ways. But given a crisis – a flood, a hurricance, and so forth – the response of the body politic is predictable. A command center is set up. The command center acknowledges that the city's normal, conditioned, reactive mechanisms are not now sufficient. The situation is unfamiliar and could be dangerous. Information must be assimilated and integrated from many sources to adjudicate difficult choices and mediate conflicting goals.

Normally, the behavioral systems of animals (water, food, travel, safety) can also operate autonomously and in their highly trained ways. But there are also times of difficulty – the water-hole dry, a predator interposed between self and young, a canopy-trail home wind damaged, a position of dominance suddenly challenged. These situations cannot be left to habit, to automaticity, or to autonomous gradients of approach and avoidance. The habits do not exist. The situation is novel and untrained. The problem may be multidimensional with difficult choices and conflicting goals. We believe that animals too would have benefited from being able to create the mind's command center for times of uncertainty and difficulty. Working consciousness is ideal for integrating multiple streams of information, for resolving conflicting goals, for coping with the novel and unfamiliar, and for maneuvering on complex optimality surfaces. Working consciousness is the perfect referee for life's close calls. Something like a working consciousness, some cognitive command center, may thus have substantial phylogenetic breadth.

Therefore, we invite colleagues to take seriously the claims of a hundred years of cognitive scientists who noted that the highest levels of information processing and particularly consciousness present themselves when difficulty, complexity, and indeterminancy are encountered. Only we suggest that this is not just a human phenomenon but rather a general functional property of adaptive minds. If you watch an aging cat consider a doubtful leap onto the dryer, you will suspect that what James (1890/1952, p. 93) said is true, "Where indecision is great, as before a dangerous leap, consciousness is agonizingly intense." All the tasks considered in this article place the human and animal participant on just the same doubtful knife-edge of decision, though in the perceptual or cognitive domain. This makes us think that these tasks are well structured and well positioned to elicit these higher-level and possibly conscious regulatory processes in animals. This, of course, is not the same thing as knowing that they do so. It remains an important goal to ask whether the present paradigms can be extended in ways that allow the stronger inference that animals are showing not just the functional features of or parallels to human conscious cognition, but its actual states and feelings.

## References

Angell, F. (1907) On judgments of "like" in discrimination experiments. *American Journal of Psychology* 18:253.

Atkinson, R. C. & Juola, J. F. (1974) Search and decision processes in recognition memory. In: *Learning, memory, and thinking*, ed. D. H. Krantz, R. C. Atkinson, R. D. Luce & P. Suppes. Freeman.

Au, W. W. & Moore, P. W. (1990) Critical ratio and critical bandwidth for the Atlantic bottlenose dolphin. *Journal of the Acoustical Society of America* 88:1635-38.

Blough, D. S. (1958) A method for obtaining psychophysical threshold from the pigeon. *Journal of the Experimental Analysis of Behavior* 1:31-43.

Boneau, C. A. & Cole, J. L. (1967) Decision theory, the pigeon, and the psychophysical function. *Psychological Review* 74:123-35.

Boring, E. G. (1920) The control of attitude in psychophysical experiments. *Psychological Review* 27:440-52.

Brown, A. S. (1991) A review of the tip-of-the-tongue experience. *Psychological Bulletin* 109:204-23.

Brown, W. (1910) The judgment of difference. *University of California Publications in Psychology* 1:1-71.

Brown, A. L., Bransford, J. D., Ferrara, R. A. & Campione, J. C. (1982) Learning, remembering, and understanding. In: *Handbook of child psychology* vol. 3, ed. J. H. Flavell & E. M. Markman. Wiley.

Byrne, R. W. & Whiten, A. (1991) Computation and mindreading in primate tactical deception. In: *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, ed. A. Whiten. Basil Blackwell.

Carter, D. E. & Werner, T. J. (1978) Complex learning and information processing by pigeons: A critical analysis. *Journal of the Experimental Analysis of Behavior* 29:565-601.

Castro, C. A. & Larsen, T. (1992) Primacy and recency effects in nonhuman primates. *Journal of Experimental Psychology: Animal Behavior Processes* 18:335-40.

Cheney, D. L. & Seyfarth, R. M. (1990) *How monkeys see the world*. University of Chicago Press.

Commons, M. L., Nevin, J. A. & Davison, M. C. eds. (1991) *Signal detection: Mechanisms, models, and applications*. Erlbaum.

Cowey, A. & Stoerig, P. (1992) Reflections on blindsight. In: *The neuropsychology of consciousness*, ed. D. A. Milner & M. D. Rugg. Academic Press.

(1995) Blindsight in monkeys. *Nature* 373:247-49.

Cumming, W. W. & Berryman, R. (1961) Some data on matching behavior in the pigeon. *Journal of the Experimental Analysis of Behavior* 4:281-84.

Davison, M., McCarthy, D. & Jensen, C. (1985) Component probability and component reinforcer rate as biasers of free-operant detection. *Journal of the Experimental Analysis of Behavior* 44:103-20.

Dewey, J. (1934/1980) *Art as experience*. Perigee Books.

Dunlosky, J. & Nelson, T. O. (1992) Importance of the kind of cue for judgments of learning (JOL) and the delayed JOL effect. *Memory & Cognition* 20:374-80.

(1997) Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language* 36:34-39.

Elliott, R. & Dolan, R. J. (1999) Differential neural responses during performance of matching and nonmatching to sample tasks at two delay intervals. *The Journal of Neuroscience* 19:5066-73.

Etkin, M. & D'Amato, M. R. (1969) Delayed matching-to-sample and short-term memory in the capuchin monkey. *Journal of Comparative & Physiological Psychology* 69:544-49.

Farthing, G. W. & Opuda, M. J. (1974) Transfer of matching-to-sample in pigeons. *Journal of the Experimental Analysis of Behavior* 21:199-213.

Fernberger, S. W. (1914) The effect of the attitude of the subject upon the measure of sensitivity. *American Journal of Psychology* 25:538-43.

(1930) The use of equality judgments in psychophysical procedures. *Psychological Review* 37:107-12.

Flavell, J. H. (1979) Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34:906-11.

Fobes, J. L. & King, J. E. (1982) *Primate behavior*. Academic Press.

Fujita, K. (1982) An analysis of stimulus control in two-color matching-to-sample behaviors of Japanese monkeys (*Macaca fuscata*). *Japanese Psychological Research* 24:124-35.

Gallup, G. G. (1982) Self-awareness and the emergence of mind in primates. *American Journal of Primatology* 2:237-48.

Gallup, G. G., Jr. & Suarez, S. D. (1986) Self-awareness and the emergence of mind in humans and other primates. In: *Psychological perspectives on the self, vol. 3*, ed. J. Suls & A. Greenwald. Erlbaum.

George, S. S. (1917) Attitude in relation to the psychophysical judgment. *American Journal of Psychology* 28:1-38.

Gilden, D. L. & Wilson, S. G. (1995) On the nature of streaks in signal-detection. *Cognitive Psychology* 28:17-64.

Gray, J. A. (1995) The contents of consciousness: A neuropsychological conjecture. *The Behavioral and Brain Sciences* 18:659-722.

Hampton, R. R. (2001) Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences* 98: 5339-62.

Hart, J. T. (1965) Memory and the feeling-of-knowing experiments. *Journal of Educational Psychology* 57:347-49.

Hays, W. L. (1981) *Statistics*. CBS College.

Herman, L. M. (1975) Interference and auditory short-term memory in the bottlenosed dolphin. *Animal Learning and Behavior* 3:43-48.

Herman, L. M. & Arbeit, W. R. (1972) Frequency difference limens in the bottlenose dolphin: 1-70 kc/s. *Journal of Auditory Research* 2:109-20.

Herrnstein, R. J. (1990) Levels of stimulus control: A functional approach. *Cognition* 37:133-66.

Heyes, C. M. (1998) Theory of mind in nonhuman primates. *The Behavioral and Brain Sciences* 21:101-48.

Holmes, P. W. (1979) Transfer of matching performance in pigeons. *Journal of the Experimental Analysis of Behavior* 31:103-14.

Humphrey, N. K. (1976) The social function of the intellect. In *Growing points in ethology*, ed. P.P.G. Bateson & R.A. Hinde. Cambridge University Press.

Inman, A. & Shettleworth, S. J. (1999) Detecting metamemory in nonverbal subjects: A test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* 25:389-95.

James, W. (1890/1952) The principles of psychology. In: *Vol. 53, Great Books of the Western World.* University of Chicago Press.

Jastrow, J. (1888) A critique of psycho-physic methods. *American Journal of Psychology* 1:271-309.

Karoly, P. (1993) Mechanisms of self-regulation: A systems view. *Annual Review of Psychology* 44:23-52.

Koriat, A. (1993) How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review* 100:609-39.

MacMillan, N. A. & Creelman, C. D. (1991) *Detection theory: A user's guide*. Cambridge University Press.

Menzel, C. R. (1999) Unprompted recall and reporting of hidden objects by a chimpanzee (*Pan troglodytes)* after extended delays. *Journal of Comparative Psychology* 113:426-34.

Metcalfe, J. & Shimamura, A. (1994) *Metacognition: Knowing about knowing*. Bradford Books.

Miller, J. T., Saunders, S. S. & Bourland, G. (1980) The role of stimulus disparity in concurrently available reinforcement schedules. *Animal Learning & Behavior* 8:635-41.

Morgan, C. L. (1906) *An introduction to comparative psychology*. Walter Scott.

Nelson, T. O. ed. (1992) *Metacognition: Core readings*. Allyn and Bacon.

Nelson, T. O. (1996) Consciousness and metacognition. *American Psychologist* 51:102-16.

Nelson, T. O. & Dunlosky, J. (1991) The delayed-JOL effect: When delaying your judgments of learning can improve the accuracy of your metacognitive monitoring. *Psychological Science* 2:267-70.

Nelson. T. O. & Narens, L. (1990) Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation* 26:125-41.

Parker, S. T., Mitchell, R. W. & Boccia, M. L. eds. (1994) *Self-awareness in animals and humans*. Cambridge University Press.

Premack, D. (1978) On the abstractness of human concepts: Why it would be difficult to talk to a pigeon. In: *Cognitive processes in animal behavior*, ed. S. H. Hulse, H. Fowler & W. K. Honig. Erlbaum.

Reder, L. M. ed. (1996) *Implicit memory and metacognition*. Erlbaum.

Roberts, W. A. & Kraemer, P. J. (1981) Recognition memory for lists of visual stimuli in monkeys and humans. *Animal Learning & Behavior* 9:587-94.

Sands, S. F. & Wright, A. A. (1980) Primate memory: Retention of serial list items by a rhesus monkey. *Science* 209:938-39.

Schull, J. & Smith, J. D. (1992) Knowing thyself, knowing the other: They're not the same. *The Behavioral and Brain Sciences* 15:166-67.

Schusterman, R. J. & Barrett, B. (1975) Detection of underwater signals by a California sea lion and a bottlenose porpoise: Variation in the payoff matrix. *Journal of the Acoustical Society of America* 57:1526-37.

Schwartz, B. L. (1994) Sources of information in metamemory: judgments of learning and feelings of knowing. *Psychonomic Bulletin and Review* 1:357-75.

Schwartz, B. L., Colon, M. R., Sanchez, I. C., Rodriguez, I. A. & Evans, S. (2002). Single-trial learning of "what" and "who" information in a gorilla (*Gorilla gorilla gorilla*): Implications for episodic memory. *Animal Cognition* 5: 85-90.

Shields, W. E. (1999) *Nonverbal judgments of knowing by humans and rhesus monkeys.* Doctoral dissertation, State University of New York at Buffalo.

Shields, W. E., Smith, J. D. & Washburn, D. A. (1997) Uncertain responses by humans and rhesus monkeys (*Macaca mulatta*) in a psychophysical same-different task. *Journal of Experimental Psychology: General* 126:147-64.

Shields, W. E., Smith, J. D., Guttmannova, K. & Washburn, D. A. (2002) A study of retrospective confidence judgments by rhesus monkeys. Manuscript in preparation.

Shiffrin, R. M. & Schneider, W. (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* 84:127-90.

Smith, J. D. (1995) The homunculus at home. Commentary on J. A. Gray, The contents of consciousness: A neuropsychological conjecture. *The Behavioral and Brain Sciences* 18:697-98.

Smith, J. D. & Minda, J. P. (1998) Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:1411-36.

(2000) Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:3-27.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R. & Erb, L. (1995) The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General* 124:391-408.

Smith, J. D., Shields, W. E., Schull, J. & Washburn, D. A. (1997) The uncertain response in humans and animals. *Cognition* 62:75-97.

Smith, J. D., Shields, W. E., Allendoerfer, K. R. & Washburn, D. A. (1998) Memory monitoring by animals and humans. *Journal of Experimental Psychology: General* 127:227-50.

Smith, J. D. & Schull, J. (1989) *A failure of uncertainty monitoring in the rat*. Unpublished data.

Smith, S. M., Brown, J. M. & Balfour, S. P. (1991) TOTimals: A controlled experimental method for studying tip-of-the-tongue states. *Bulletin of the Psychonomic Society* 29:445-47.

Swartz, K. B., Sarauw, D. & Evans, S. (1999) Comparative aspects of mirror self-recognition in great apes. In: *The mentalities of gorillas and orangutans: Comparative perspectives*, ed. S. T. Parker, R. W. Mitchell & M. L. Boccia. Cambridge University Press.

Swartz, K. B. (1997) What is mirror self-recognition in nonhuman primates, and what is it not? In: *The self across psychology: Self-recognition, self-awareness, and the self concept. Annals of the New York Academy of Sciences*, ed. J. G. Snodgrass & R. L. Thompson 818:65-71. New York Academy of Sciences.

Swets, J. A., Tanner, W. P. & Birdsall, T. G. (1961) Decision processes in perception. *Psychological Review* 68:301-40.

Teller, S. A. (1989) *Metamemory in the pigeon: Prediction of performance on a delayed matching to sample task*. Undergraduate thesis, Reed College.

Terman, M. & Terman, J. (1972) Concurrent variation of response bias and sensitivity in an operant-psychophysical test. *Perception and Psychophysics* 11:428-32.

Thomson, G. H. (1920) A new point of view in the interpretation of threshold measurements in psychophysics. *Psychological Review* 27:300-07.

Tolman, E. C. (1927) A behaviorist's definition of consciousness. *Psychological Review* 34:433-39.

(1932/1967) *Purposive behavior in animals and men.* The Century Company.

(1938) The determiners of behavior at a choice point. *Psychological Review* 45:1-41.

Treisman, M. & Faulkner, A. (1984) The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance* 10:119-39.

Urban, F. M. (1910) The method of constant stimuli and its generalizations. *Psychological Review* 17:229-59.

Washburn, D. A. & Rumbaugh, D. M. (1992) Testing primates with joystick-based automated apparatus: Lessons from the Language Research Center's Computerized Test System. *Behavior Research Methods, Instruments, and Computers* 24:157-64.

Washburn, D. A., Smith, J. D., Baker, L. A. & Raby, P. R. (2001) Responding to uncertainty: Individual differences and training effects. *Proceedings of the 2001 meeting of the Human Factors and Ergonomics Society*, 911-15.

Watson, C. S., Kellogg, S. C., Kawanishi, D. T. & Lucas, P. A. (1973) The uncertain response in detection-oriented psychophysics. *Journal of Experimental Psychology* 99:180-85.

Weiskrantz, L. (1986) *Blindsight: A case study and implications*. Oxford University Press.

(1997) *Consciousness lost and found: A neuropsychological exploration*. Oxford University Press.

Whiten, A. & Byrne, R. W., eds. (1997) *Machiavellian intelligence II: Extensions and evaluations*. Cambridge University Press.

Woodworth, R. S. (1938) *Experimental psychology*. Holt.

Wright, A. A., Santiago, H. C., Sands, S. F., Kendrick, D. F. & Cook, R. G. (1985) Memory processing of serial lists by pigeons, monkeys, and people. *Science* 229:287-89.

Wright, A. A., Shyan, M. R. & Jitsumori, M. (1990) Auditory same/different concept learning by monkeys. *Animal Learning & Behavior* 18:287-94.

Yunker, M. P. & Herman, L. M. (1974) Discrimination of auditory temporal differences by the bottlenose dolphin and by the human. *Journal of the Acoustical Society of America* 56:1870-75.

## Appendix 1: Details of simulations and formal models

**The same-different task of Shields et al. (1997)**. Figure 6b showed the results when animals were presented with 13 trial types. Remember that each trial presented the animal with two boxes. If both had Density Level 13, the disparity between them was 0.0 and the trial was Same. As the two boxes had Level 12 and 13 density, or Level 11 and 13 density, and so forth, the trial was an easier and easier Different. One step in modeling was to psychologically scale the objective density-disparity continuum into the subjective-impression continuum that described best animals' perceptual sensitivities in the task. This scaling was done as follows. First, we took the natural logarithm of the 13 relevant pixel densities so that the continuum would obey Weber's law (with the same proportional change in density creating the same psychological change in density). Second, we found the positively signed difference in logarithmic density between each level of disparity and Level 13 (Same). This subtraction translated the scale so that Same trials had a value of 0.0 on the disparity continuum (Level 13 - Level 13), whereas progressively easier Different trials (Level 12 and 13, Level 11 and 13, etc.) had larger and larger positive values. Third, these logarithmic difference values were rescaled through being multiplied by a free parameter (Stretch) in the model to the point that the separations of the 13 values on the continuum correctly reflected animals' sensitivities (assuming, as SDT does, that perceptual error causes subjective impressions of disparity to scatter in unit-normal distributions [SD=1.0] around the objective or expected value of the stimulus event). The stretched scale ran from 0.0 out to 5.913 for the largest disparity presented.

The second step in modeling was to place the Different-Uncertain and Uncertain-Same criteria along this disparity continuum in the way that would most closely mirror the animals' decisional strategy. We searched for these best-fitting placements as follows. We sampled the Different-Uncertain criterion placement at every 1% of the range of the underlying disparity continuum (from 0.0 out to 5.913). We sampled the Uncertain-Same criterion placement at every 1% of the range of the underlying continuum from 0.0 up to the level of the Different-Uncertain criterion placement then in force. That these two criterion points would reverse positions along the disparity continuum makes no sense.

Combining these two steps of modeling, the complete search for the best-fitting parameter configuration of the SDT model evaluated 101 Stretch values, 101 Different-Uncertain criterion placements, and varying numbers of Uncertain-Same criterion placements for a total of 520,251 simulants who each completed 8,000 trials in a virtual version of the monkeys' same-different discrimination. On each trial, the simulant received a trial at 1 of 13 disparity levels, misperceived the disparity according to the perceptual error assumed in SDT, and responded according to its two criterion placements. Summarizing the 8,000 trials, we virtually drew the graph of the simulant's performance profile and compared it mathematically to the performance profile shown in Figure 6b. The criterion of best fit was the sum of the squared deviations (SSD) between the 39 observed and simulated response percentages. The value of SSD for the best-fitting simulant was 810 (39 deviations of around 4% – squared then summed – produced this total). We also calculated the value of a more intuitive fit index, the average absolute deviation (AAD), which is the average amount the simulated response percentages deviate from those observed. The best-fitting AAD was 3.8%. The best-fitting parameter values were 7.07 (Stretch), 0.710 (Same-Uncertain criterion), and 1.360 (Uncertain-Different criterion).

**Optimality in the same-different task of Shields et al. (1997).** We retained the scaling of the logarithmic disparity axis that fit best animals' sensitivities. These sensitivities define the animals' perceptual limits that cannot be increased in the service of increasing rewards. Along the disparity axis, we surveyed the reward efficiency of strategies that placed the center of the Uncertain response region at 101 places at each 1% increment along the range of the disparity continuum. At each center, we examined the reward efficiency of strategies that gradually widened out the Uncertain response region from having 0 width (zero 1% increments to either side of center) up to 50 width (fifty 1% increments to either side of center). These 5,151 simulants each received 8,000 trials in the virtual SD task, once again receiving 1 of 13 disparities, misperceiving that disparity, and responding according to their two criterion points. These simulants were also subject to the trial times and penalty times of the actual task, so that we could estimate the rewards per minute that each decisional strategy would have received in the actual task and assess the optimality of the monkeys' actual decisional strategy.

**The SPR task of Smith et al. (1998).** Figure 7a showed the results when animals were presented with 5 trial types – there were probes that were not in the list and probes that repeated List Items 1, 2, 3, or 4. We assumed that probes on Not There trials contacted on average but with memory variability traces of strength 0.0 (SD=1.0; the leftmost normal curve in Figure 10a), whereas probes on There trials contacted on average but with memory variability trace strengths at the memory sensitivity (d') appropriate to the performance that the animal showed at each serial position (SD=1.0; the four rightward normal curves in Figure 10a; see MacMillan & Creelman 1991, pp. 209-30). Thus probes of the animal's better (more sensitive) serial positions would encounter stronger traces that lie on average farther from zero on the trace-strength continuum. The first step of the modeling – the scaling of the underlying subjective-impression axis – is given by the animal's sensitivities (d's) so there is no need to stretch or scale the continuum.

Given this underlying continuum, for the forced condition that disallowed the Uncertain response, we evaluated the position of one criterion parameter separating the Not There and There response regions because these were the only two responses granted the animal in that condition. For the optional condition that allowed the Uncertain response, we evaluated the position of two criterion parameters establishing the Not There, Uncertain, and There response regions. The fitting procedure evaluated 61 levels of each criterion point for a total of 226,981 simulants who each completed 8,000 trials in a virtual version of the SPR task. On each trial, the simulant received one of 5 trial types (Not There or a probe of one of four serial positions), assessed (with memory variability) the trace strength this probe item contacted, and responded according to its single or twin criterion placements. Once again we summarized over 8,000 trials the simulant's performance profile and compared it mathematically to the animal's observed performance profile. The criterion of best fit was the sum of the squared deviations (SSD) between the 15 observed and simulated response percentages. The value of SSD for the best-fitting simulant was 210 (15 deviations of about 3% – squared then summed – produced this total). The value of the intuitive fit index, the average absolute deviation (AAD), was 3.1%.

**Optimality in the SPR task of Smith et al. (1998).** We retained the trace-strength continuum that ran from zero up to higher d's for the animal's more sensitive serial positions, assuming that memory sensitivities were also an information-processing limit that could not be increased in the service of greater rewards. Focusing on the condition with the Uncertain response allowed, we surveyed the reward efficiency of strategies that placed the center of the Uncertain response region at 101 places at each 1% increment along the trace-strength continuum, and, given each center, that widened the Uncertain response region out from having 0 width (zero 1% increments to either side of center) up to 50 width (fifty 1% increments to either side of center). These 5,151 simulants each received 8,000 trials in the virtual SPR task, subject to the trial times, penalty times, and reward structure of the actual task, and responding in accordance with the three response regions in effect.

**Monkey 1's performance in the DMTS task of Hampton (2001).** We assumed that samples left behind memory impressions at the d' (SD=1.0) along the trace-strength continuum appropriate for each delay condition. We assumed that the three foils presented in the memory test were associated with average trace strengths of 0.0 (SD=1.0). Note that memory variability on a trial could cause the sample's trace to be less active than a foil's trace, causing errors in the task, and causing more errors in the task for longer delays with their lower d's. On forced trials, no criterion point applies because the animal must complete all memory tests and it is only a matter of whether he is correct. On optional trials, one criterion line separates the Decline and Accept response regions. The placement of this criterion is the only parameter in this SDT model because here too the scaling of the underlying representational axis is fixed by the d's. The fitting procedure for Monkey 1 evaluated 201 levels of the Decline-Accept criterion, for a total of 201 simulants that each completed 8,000 trials in the virtual DMTS task. The values of SSD and AAD for the best-fitting simulant were, respectively, 320 and 3.7%.

Monkey 2's performance in the DMTS task of Hampton (2001). We modeled the performance of Hampton's Monkey 2 using different procedures that focus on the character of performance when the metacognitive data pattern is not clearly seen. In this case, we assumed that long delays became associated with errors and timeouts and with a reduced urge to complete memory tests. By using this temporal cue, the animal could decline memory tests adaptively at delays, as Monkey 2 and Teller's pigeons did, without consulting any memory trace at all. We assumed that the animal began every trial with a level of Trial-Accept Determination that decayed exponentially as the delay interval transpired. We set this Accept Determination at 3.0 and then we searched a free parameter (Decay) to estimate the steepness of the loss of courage (determination) as time went by. We assumed that at the choice point, the animal's remaining Accept Determination was scattered (SD=1.0 across trials) and that he placed a criterion point on the Accept Determination continuum that let him decline the trials with low remaining determination and accept the trials with high remaining determination. The target data were Monkey 2's 12 data points (Table 1, Row 2). The parameter search for this model involved evaluating two parameters – Decay and the placement of the Decline-Accept Criterion.

The fitting procedure evaluated 21 levels of Decay and 201 levels of the Decline-Accept Criterion, for a total of 4,221 simulants that each completed 8,000 trials in the virtual DMTS task. The best-fitting values of the model were 0.994 (per second) for the decay, and a criterion setting of 1.64 along the Accept-Determination continuum. The values of SSD and AAD for the best-fitting simulant were, respectively, 90 and 2.1%. Table 1 (Row 5) shows that this simulant reproduced well the performance of Monkey 2 (Row 2). The closeness of this fit shows that Monkey 2's performance was essentially not metacognitive, though he may have used a tiny amount of memory monitoring. In fact, when we fit this animal's data making metacognitive assumptions instead, we found a best-fit index that was 12 times as high (i.e., 12 times worse) than when we made the non-metacognitive assumption. The two monkeys in Hampton's experiment showed a clear and interesting individual difference, and illustrate two different kinds of performance in the DMTS task. We have already seen the data pattern of Monkey 2 essentially replicated with pigeons (Teller 1989), and would anticipate similar results if this model were applied to those data.

## Author Notes

J. David Smith, Department of Psychology and Center for Cognitive Science, State University of New York at Buffalo.

Wendy E. Shields, Department of Psychology, University of Montana.

David A. Washburn, Department of Psychology and Language Research Center, Georgia State University.

Correspondence concerning this article should be addressed to J. David Smith, Department of Psychology, Park Hall, State University of New York at Buffalo, Buffalo, NY, 14260, or to psysmith@acsu.buffalo.edu.

The preparation of this article was supported by the National Institutes of Health (HD-38051).