Psychonomic Bulletin & Review
1998, 5 (3), 370-389

Understanding variability in binary and continuous choice

DANIEL FRIEDMAN and DOMINIC W. MASSARO
University of California, Santa Cruz, California

Excessive variability in binary choice (categorical judgment) can take the form of probability matching rather than the normatively correct behavior of deterministically choosing the more likely alternative. Excessive variability in continuous choice (judgment rating) can take the form of underconfidence, understating the probability of highly likely events and overstating the probability of very unlikely events. We investigated the origins of choice variability in terms of noise prior to decision (at the evidence stage) and at the decision stage. A version of the well-known medical diagnosis task was conducted with binary and continuous choice on each trial. Noise at evidence stage was reduced by allowing the subjects to view historical summaries of prior relevant trials, and noise at the decision stage was reduced by giving the subjects a numerical score on the basis of their continuous choice and the actual outcome. Both treatments greatly reduced variability. Cash payments based on the numerical score had a less reliable incremental effect in our experiment. The overall results are more consistent with a Logit model of decision than with a simple criterion (or maximization) rule or a simple probability-matching rule.

One of the most striking aspects of human behavior is its variability. Sometimes a poker player might open with a small pair; other times he might check. In the laboratory, a given subject presented with a given description of a hypothetical person (*entertaining, vain, and agreeable*, for example) might judge the hypothetical person to be introverted in some trials and extroverted in other trials. Townsend and Busemeyer (1995), for instance, reported considerable variation across trials in the cash equivalent estimated by a given subject for a given gamble.

Variability might be expected when there is no objectively correct answer or when a random response is optimal. However, it also occurs in situations in which a specific invariant behavior is objectively better. Subjects are asked, for example, to predict which of two lights will flash, given experience that the left one flashes about 70% of the time. To maximize the probability of correct prediction, subjects should *always* choose the left light, but typically they choose the right light about 30% of the time. This sort of variability is called *probability-matching* behavior. Such behavior has been observed in many probability-learning experiments in humans and other animals (Davison & McCarthy, 1988; Estes, 1954, 1984; Myers, 1976; Thomas & Legge, 1970). Although overshooting is sometimes observed (Massaro, 1969; Myers & Cruse, 1968), people clearly are not responding with the more frequent alternative all of the time.

Another sort of variability has also been documented. When asked to estimate the likelihood of very improbable or very probable events (say, objective probability .01 or .99), subjects often report (or act as if they use) less decisive subjective probabilities (say, subjective probability of .15 or .85). This sort of variability, which we refer to as *undershooting*, has been interpreted as conservatism, or underconfidence (Camerer, 1995; Kahneman & Tversky, 1979).

Faced with such variability, theorists have found it difficult to provide a coherent description of human decision making. The general goal of the present paper is to provide a framework with which to measure and to identify the processes that lead to the observed variability. A simple conceptual model frames our inquiry. As is shown in Figure 1, we decompose the choice process in any given task into two stages—evidence and decision. The input to the evidence stage is all stimuli (both internal and external to the subject) associated with the task, and the output is a subjective value $g(r)$ for each response alternative r available in the task. The value $g(r)$, which we refer to as the goodness of the response alternative, may accurately reflect the objective evidence, may be biased in some way, or may incorporate noise to some degree (Green & Swets, 1966; McClelland, 1991). Given the $g(r)$ for each possible choice r, the decision stage selects a specific response r^*, and this selection process may itself be noisy to some degree.

This work was supported by NSF Grant SBR 9310347 and Public Health Service Grant PHS R01 DC 00236 and was presented at the 1996 Economic Science Association meeting. We are grateful to Stephen Kitzis, Tamara Torrence, Anne Shin, and Lisa Lima for research assistance on earlier stages of the project and to Nitai Farmer, Eric Berg, and Hugh Kelley for helping us run the main set of experiments and analyze the data reported here. Hugh Kelley and Seble Menkir polished the tables and figures. The final exposition owes much to Duncan Luce, Gregg Oden, an anonymous referee, and Richard Schweickert. D.F. is in the Department of Economics. Correspondence concerning this article should be sent to D. W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

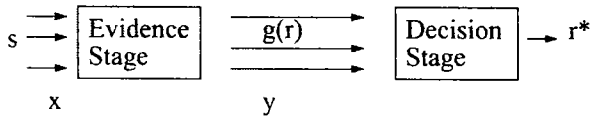


Figure 1. Evidence and decision stages in any given task. The input to the evidence stage is all stimuli (both internal and external to the subject) associated with the task, and the output is a subjective value $g(r)$ for each response alternative r available in the task. The *goodness* values $g(r)$ are inputs to the decision stage, whose output is a specific response r^*.

In this paper, we analyze formally both origins of variability and assess them empirically. The theoretical analysis develops three alternative rules governing the decision stage: the simple maximization or criterion rule, the standard stochastic decision rule of matching estimated probabilities, and a flexible stochastic rule known as Logit. The rules offer differing predictions on how observed responses depend on noise at evidence and at decision.

The empirical work involved only specific variants of the well-known medical diagnosis task (Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Friedman, Massaro, Kitzis, & Cohen, 1995; Gluck & Bower, 1988; Nosofsky, Kruschke, & McKinley, 1992). Each trial, the subject viewed a *medical chart* (a list of symptoms; each was binary valued in most previous work, but here each symptom had four possible values), chose a *diagnosis*, and then was told the actual disease, which was stochastically related to the symptoms. We used two response modes for reporting the diagnosis—binary and continuous. In the binary mode (also known as categorical choice), the subjects chose either disease A or disease B as the more likely. In the continuous mode (also known as ratings judgment), the subjects reported the degree of confidence in their choice of A or B.

There is a history of both binary choice and continuous judgments in experimental psychology, psychophysics, and judgment. Continuous judgments are employed in magnitude estimation tasks (Marks, 1976) and information integration experiments (N. H. Anderson, 1981, 1982) and can even be considered to be reflected in the reaction times for making discrete judgments (Luce, 1986). In the medical diagnosis task, on the other hand, psychologists routinely analyze binary choice data (usually after averaging across subjects and trials) but not continuous choice data. For example, Nosofsky et al. (1992) collected both sorts of data but neglected the continuous choices in comparing competing models. They correctly observed that "there is no generally agreed-on method for using the models to predict direct probability estimates" (p. 219). A secondary goal of our paper is to show that continuous choice data (including direct probability estimates) contain potentially valuable fine-grained information on the level of individual subjects (Friedman et al., 1995) and to show how to extract that information even when there is variability associated with the rating judgments, as might be found with underconfidence.

The innovative work of McDowell and Oden (1995) substantiated the value of utilizing both binary and continuous dependent measures and highlighted several important issues for the current investigation. Their subjects were presented with cursive test words that varied in ambiguity between the target words *eat* and *lot*. Most sessions required both response modes—binary (which target word is closest to the test word?) and continuous (to what degree?). Separate sessions with a single response mode indicated that responses in either mode were not systematically affected by the presence of the second mode. Their most striking finding was that the relation between the two modes was monotone but nonlinear. Small changes in the stimulus word produced larger changes in the continuous response than in the average binary response for the relatively unambiguous test words, and the reverse was true for the more ambiguous words. That is, in terminology explained below, the binary response overshoots relative to the continuous response. We shall explore the interpretation that binary categorization necessarily loses information about what the subject perceives (and knows) and that the information tends to be preserved in the continuous rating judgment. In both types of judgment, however, the processes involved in the task can be made more transparent by accounting for their variability.

Our research strategy was to manipulate laboratory conditions that affect the amount of noise at the evidence stage and the amount of noise at the decision stage. Noise at the evidence stage was varied by providing (or not providing) summaries of the association between medical charts and diseases seen in previous trials. Noise at the decision stage was varied by providing trial-by-trial feedback only on the actual disease, by also providing a continuous score based on the response and the actual disease, or by also offering a substantial cash payment on the basis of the score. The data from the two response modes allow us to infer the relative contribution of the two sources of variability for individual subjects.

The next section presents a simple two-stage information processing framework and the three alternative models of the second (decision) stage. It summarizes the models' distinctive predictions regarding the variability of binary and continuous choice and the impact of noise at each stage. It also offers a new theoretical explanation of underconfidence, based on a selection bias when evidence is noisy. The underlying mathematical arguments are collected in an Appendix.

The rest of the paper is empirical. We describe the medical diagnosis experiment and the data and construct two quantitative measures of choice variability. The measures (called *overshooting* and b, a regression slope coefficient) clearly indicate variability for both choice modes in the baseline conditions. With the partial exception of cash payments, the treatments intended to reduce variability actually did reduce (in some cases, virtually eliminate) choice variability according to our measures. Overall, the flexible decision rule (Logit) predicted the observed

choice data better than did the alternatives we consider. In the last section, we offer some interpretations and suggestions for further work.

The Evidence Stage

The task defines the objective stimuli $s \in J$ and the available responses $r \in I$. We assume that there is a sufficient statistic, $x = x(s)$, that would objectively code each stimulus s in the absence of noise. When noise is present at the evidence stage, the subjects do not have access to x but only to a noisy version of it, $y = x + e$, where the error or noise, e, has mean and mode zero (Green & Swets, 1966; McClelland, 1991). In general, x and y are vectors, but for the task considered in this paper, both are scalars in the unit interval [0,1]. In our medical diagnosis task, $x(s)$ is the true posterior probability of disease A, given symptom configuration s, and y is the corresponding subjective estimate. In the letter identification task of Massaro and Hary (1986), $x(s)$ summarized the objective evidence for one target letter in the test letter s, and y was the corresponding subjective estimate.

Recall that, in our information processing framework, the evidence stage transmits to the decision stage a scale value $g(r)$, the *goodness*, for each available response r. By assumption, y is a sufficient statistic for the noisy evidence, so there is a function that expresses the $g(r)$ in terms of y. That is, there is some function G such that $g(r) = G(r,y)$ for each available response $r \in I$. Examples of goodness functions G will be presented below.

We should note that it is possible to model the evidence stage in more detail. Massaro and Friedman (1990), for example, decomposed it into (1) evaluation of the sources of available information and (2) integration of these sources. Our focus here is on the decision stage, so we simply summarize all the processes prior to decision as the evidence stage and summarize in e all noise introduced prior to decision. However, we should acknowledge that systematic manipulation of the separate processes contributing to the evidence outcome potentially can shed further light on decision processes.

The Decision Stage

The decision stage produces the actual response $r^* \in I$ from the goodnesses according to some general rule. For example, if the response mode is binary, with the two choices coded as 1 for disease A and 0 for disease B, the decision rule assigns either $r^* = 1$ or $r^* = 0$, given the realized scale values $g(0)$ and $g(1)$. If the response mode is continuous, with responses scaled to the unit interval $I = [0,1]$, the decision rule assigns some number $r^* \in [0,1]$, given the realized scale values $\{g(r): r \in I\}$.

Because human choices are always variable to some degree, every sensible empirical model must introduce variability at some point. Some models have variability only at the evidence stage, prior to decision (Green & Swets, 1966; McClelland, 1991), whereas other models allow variability only at the final decision stage (Massaro & Friedman, 1990). The three decision rules we consider differ primarily in whether and how they introduce variability at the decision stage.

Criterion Rule. Most economists and some psychologists favor the rule $r^* = argmax_{r \in I} g(r)$—that is, always pick a response with maximum goodness. For logical completeness, we assume that all responses with maximal $g(r)$ are equally likely but we note that ties have probability zero for standard specifications of noise at the evidence stage. Psychologists generally refer to this maximization rule as the *criterion rule*, or CR (Green & Swets, 1966; Macmillan & Creelman, 1991). With the negligible exception of resolving ties, CR is deterministic; the observed choice variability is attributed entirely to the error e introduced at the evidence stage.

The formalization of the CR originated in Thurstone (1927) case V but actually can be traced back to the assumption of a deterministic decision process in early psychophysics (Massaro, 1989, chaps. 10 and 11). The idea is that the decision process maintains a criterion or threshold level y_c in a binary choice task. The decision is based on where the noisy evidence summary y falls relative to the criterion value. If the evidence value falls on one side of the criterion, one response is made; if it falls on the other side of the criterion, the other response is made. Under the standard assumptions of signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 1991), this rule is equivalent to maximizing $g(r)$ when the criterion y_c is chosen to equalize the discriminal processes $G(1,y)$ and $G(0,y)$.

Relative Goodness Rule. Many psychologists favor a stochastic rule called the *relative goodness rule*, or RGR, also known as Luce's choice rule (Luce, 1959; Shepard, 1957). It assigns probability $p^{RGR}(r^*) = g(r^*)/\sum_{r \in I} g(r)$ to choice r^*; that is, the choice probability directly reflects the relative evidence. The probability that $r^* = 1$ is chosen in binary mode is $p^{RGR} = g(1)/[g(1) + g(0)]$. The probability density that r^* is selected in continuous mode is $f^{RGR}(r^*) = g(r^*)/\int_0^1 g(r)\,dr$.

The probability-matching behavior predicted by the RGR is nonoptimal in the sense that it does not maximize the likelihood of being correct. Consider, for example, binary choice when disease A occurs 85% of the time, and suppose that the goodnesses reflect this fact—for example, $g(1) = .85$ and $g(0) = .15$. Then, the subject responds A with probability .85 and is correct in 74.5% of the trials ($= .85^2$ correct A choices + $.15^2$ correct B choices). By contrast, the CR in this case would always select disease A and thus be correct in 85% of the trials. When viewed in another light, however, behavior following the RGR might be considered reasonable (see Friedman et al., 1995). For example, the subject may be communicating information about the rate of occurrence of some event by matching it with the rate of responding, or the subject may find it too boring to always make the same choice.

Logit Rule. Many empirical social scientists (e.g., S. Anderson, de Palma, & Thisse, 1992; Greene, 1990;

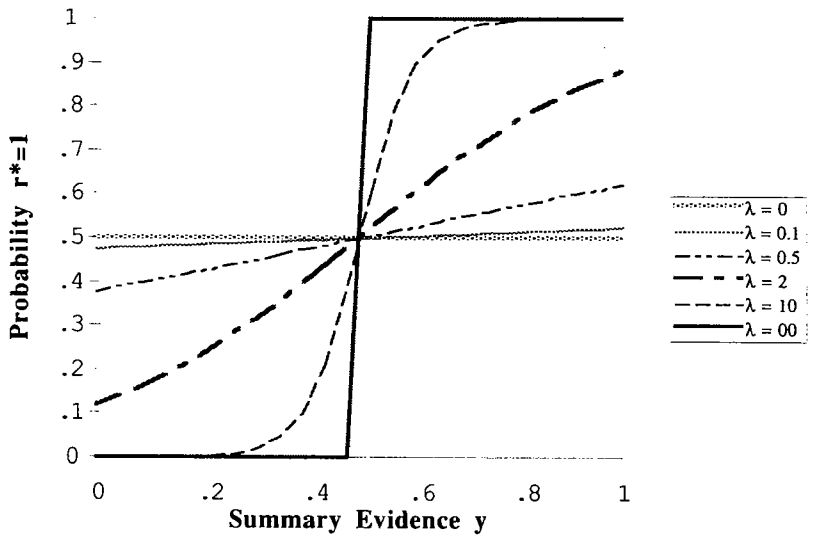


Figure 2. The probability of $r^* = 1$ in binary choice as a function of the summary evidence y according to the Logit model with parameter λ.

McKelvey and Palfrey, 1995) favor the *Logit* rule, which assigns to choice r^* the probability

$$p^{\text{Logit}}(r^* \mid \lambda) = \exp(\lambda g(r^*)) / \Sigma_{r \in I} \exp(\lambda g(r)). \quad (1)$$

One justification for the Logit rule is the roughly logarithmic relationship between many dimensions of the physical world and our psychological experience of that world. It is well known, for example, that perceived differences are a constant proportion of a light's intensity rather than an absolute difference. The name Logit is standard and comes from the fact that the log odds $\ln[p^{\text{Logit}}(r \mid \lambda)/p^{\text{Logit}}(s \mid \lambda)]$ are linear, indeed equal to λ times the difference goodness $g(r) - g(s)$. The precision parameter λ reflects the variability level in the evidence and decision processes. Lemma 2 in the Appendix shows that Logit choice probabilities converge to the noiseless CR choice probabilities as $\lambda \to \infty$, that Logit choice probabilities are noisier than RGR choice probabilities for λ sufficiently small, and that Logit choice probabilities converge to the uniform distribution as $\lambda \to 0$. The uniform distribution assigns equal probability to each possible response, independent of the evidence and goodnesses; it represents the noisiest possible decision process. Figure 2 graphs the Logit probability of $r^* = 1$ in binary choice as a function of the summary evidence y for various values of the precision parameter λ.

The Logit rule has two important advantages. First, it is quite flexible regarding the noise source and level. The CR assumes no noise at the decision stage, and the RGR necessarily predicts a specific level of decision noise—for example, binomial sampling variance for binary choice (Massaro, 1998). By contrast, Logit allows us to estimate empirically the noise level from the data via the parameter λ. Second, the Logit rule (like the CR) allows us to work with negative as well as positive scale values for $g(r)$. The RGR requires that some $g(r)$s are positive and none are negative; otherwise, it can produce nonsensical negative probabilities or probabilities exceeding 1.0.

The Logit rule is closely related to the CR. McFadden (1973), Yellott (1977), and others have shown that the CR applied to goodnesses $g(r)$ perturbed by a noise term with the extreme value (also known as a double negative exponential or Weibull) distribution will reproduce the Logit choice probabilites. The Logit rule also is related to the RGR; indeed, it is the precisely the RGR formula applied to $g(r)$s transformed with the standard exponential family $T(x; \lambda) = \exp(\lambda x)$. As was mentioned earlier, the psychophysical literature offers a justification of this transformation.

Three other related models are worth noting briefly. McDowell and Oden (1995), Tang (1996), and others have used the variable power RGR, whose choice probabilities are $g(r^*)^\lambda / \Sigma_{r \in I} g(r)^\lambda$. The power RGR, like the regular ($\lambda = 1$) RGR, requires positive goodness values for $g(r)$. When goodnesses are positive, one can take logs and put these into the Logit formula. The result is precisely the power RGR. A second closely related model is obtained by transforming arbitrary positive $g(r)$s so that they are uniformly bounded, using the transformation $U(x) = x/(1 + x)$. The composition $S = U \circ T$ is an order-preserving map of an arbitrary real number x into the point $S(x; \lambda) = \exp(\lambda x)/(1 + \exp(\lambda x))$ The connectionist (or CMP; Gluck & Bower, 1988) decision rule uses this logistic (or sigmoid or squashing) transformation S instead of the transformation T in the RGR. A third choice model, known as Probit, is also quite similar to Logit but uses the cumulative normal distribution function instead of T (see, e.g., Cheung & Friedman, 1997). Every empirical study we

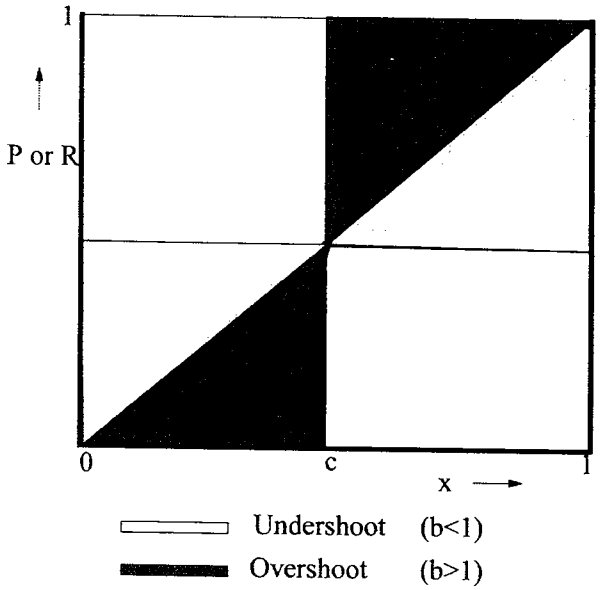


Figure 3. Undershooting and overshooting classifications for possible scatterplots of the data *P* or *R* against the true values *x*.

know that compares the Logit, Probit, power RGR, and sigmoid rules concludes that they all produce quite similar fits and inferences.

Theoretical Predictions for Binary Choice

Consider a binary choice task with varying stimuli s, and recall that $x(s) \in [0,1]$ in this context refers to the true posterior probability that $r = 1$ is better, given a particular stimulus s. Thus $x = 0$ means that $r = 0$ is certainly better, given s, whereas $x = 1$ means that $r = 1$ is certainly better, and $x = .6$ means that $r = 1$ is better with objective probability .6, and so forth. Data gathered in such binary choice tasks typically are analyzed by calculating $P(s)$, the actual proportion of $r = 1$ choices in trials with some given stimulus s. The experimenter compares $P(s)$ to the predictions of various theoretical models and perhaps to $x(s)$.

Figure 3 classifies possible scatterplots of the data $P(s)$ against the true values $x(s)$ as the stimulus s varies. Psychologists' usual null hypothesis is that the data lie near the diagonal $P = x$—that is, subjects *probability match*. We say that a subject *overshoots* if, for some critical value c near .5, the data P tend to lie below (above) the diagonal when $x < c$ (when $x > c$) and when x is not very close to 0 or 1. The qualification about the endpoints is due to the fact that there is little room for P to lie below the diagonal when x is near 0 or above the diagonal when x is near 1. If two subjects both overshoot, we will say the first overshoots more if his or her choice data are further away from the diagonal. Economists' usual null hypothesis is extreme overshooting, with $P = 0$ for $x < .5$ and $P = 1$ for $x > .5$—that is, the unit step function at one-half. It is easy to see that this pattern arises from the CR with no noise at the evidence stage (Massaro, 1987, p. 116).

Another possible choice pattern is *undershooting*, where P tends to lie below (above) the diagonal when $x > c$ (when $x < c$); see also Figure 4 of Kahneman and Tversky (1979). An extreme case is constant $P(s) = c$, unresponsive to the evidence s. For example, binary choice under the uniform distribution is constant with $c = .5$. Finally, we can have a general bias toward $r = 1$ (or toward $r = 0$) if P tends everywhere to lie above (or below) the diagonal.

In the Appendix, we show that the different models of the decision stage imply distinct predictions regarding binary choice patterns. The CR predicts that overshooting decreases as the noise at the evidence stage increases. The intuition is that noise decreases the probability that subjects correctly perceive the better alternative (especially on close calls), so P increases (decreases) for x below (above) .5 (especially on close calls). The RGR predicts choice probabilities precisely on the diagonal—probability matching with no over- or undershooting—when the evidence stage is noiseless and predicts increasingly strong undershooting as the noise at the evidence stage increases. The intuition here is that noise is more likely to increase than to decrease evidence for $x(s)$ below .5 and that the reverse is true for $x(s)$ above .5. Hence P increases for x below .5, and P decreases for x above .5. Note that the CR assumes no variability at decision, and the RGR predicts sampling variability, so neither predicts an impact for noise at the decision stage. For precision $\lambda > 2.0$, the Logit model predicts overshooting when the evidence stage is noiseless and predicts less overshooting (or more undershooting) with increasing noise at either the evidence or the decision stage.

Predictions for Continuous Choice

Some laboratory tasks offer subjects an essentially continuous choice from a range of alternatives (N. H. Anderson, 1981, 1982; Varey, Mellers, & Birnbaum, 1990). As a practical matter, choice is bounded above and below, so, perhaps after linear rescaling, the range of choices can be coded as the unit interval $I = [0,1]$. The natural way to analyze data gathered in such a continuous choice task is to calculate $R(s)$, the mean choice for a given stimulus s by a given subject (or group of subjects). The experimenter can compare $R(s)$ to the objective value $x(s)$ and to predictions of appropriate theoretical models. Again, the relation between R and x can involve over- or undershooting, depending on the the subject's decision rule.

Recall that undershooting in continuous choice can arise from underconfidence. In some contexts, underconfidence might be rationalized as arising from risk aversion or, more generally, from an asymmetric utility or loss function. For example, due to competition for market share, established professional forecasters suffer excessive penalties from forthright but wrong forecasts and so rationally will avoid forthright forecasts (see, e.g., Friedman, 1983). But such explanations seem implausible in typical repetitive laboratory tasks, where the stakes are small or nil.

Here, we offer a new explanation of undershooting and underconfidence on the basis of a selection bias. The intuition is that noisy evidence that points towards an extreme action is (in many contexts) more likely to contain errors of exaggeration than errors of moderation. There is

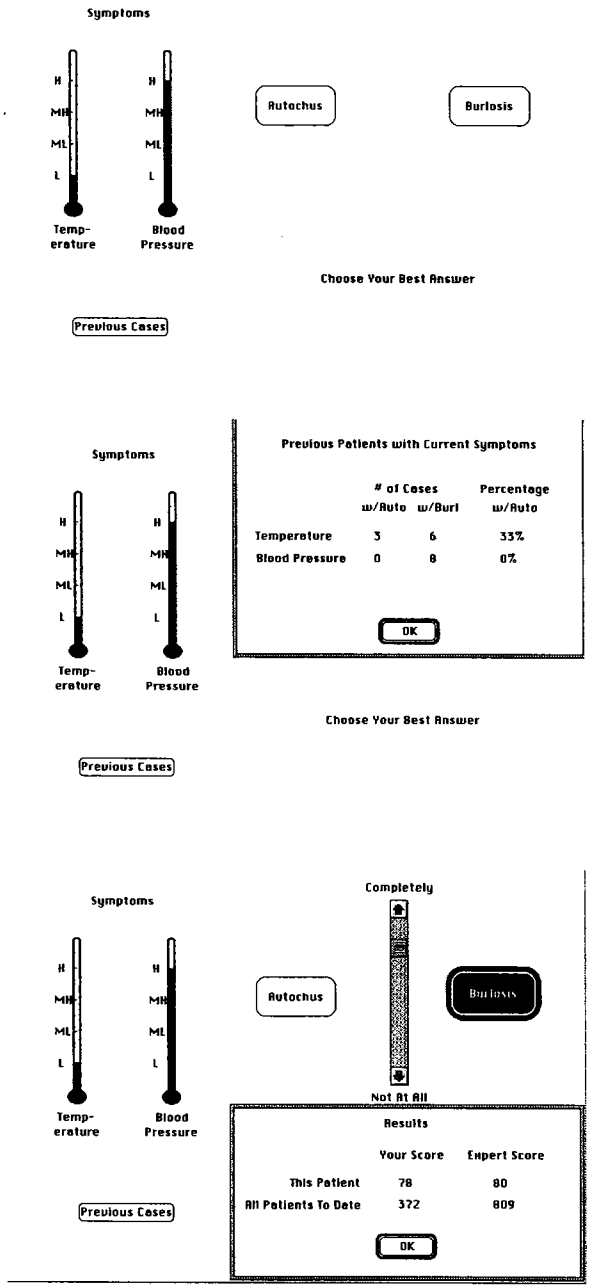


Figure 4. Examples of medical diagnosis screens used in the experiment. The three panels show, respectively, the basic screen (top), the history condition (middle), and the score (bottom).

never a *sure thing*. Indeed, if the evidence points to one of the most extreme actions (0 or 1), it could not possibly contain a significant error of moderation. Hence, evidence should be discounted more heavily when it is noisy and extreme, resulting in underconfidence and undershooting. Lemma 1 and Proposition 3 in the Appendix formalize this explanation.

One technical point remains before proceeding with the models' predictions for continuous choice. Since the true value x, the noisy signal y, and the allowable choices r are all on the same scale $I = [0,1]$, it is natural to assign goodnesses on the basis of the closeness of r to y, using, say, the squared Euclidean distance. Then, the maximum value A of closeness is attained when $r = y$, and it exceeds the minimum value (by some amount $B > 0$) when r and y are as far apart as possible. These conventions imply that the goodness function is $g(r) = G(r,y) = A - B(r - y)^2$. The Appendix shows that this natural goodness function is formally identical to the quadratic scoring rule, which has a long history in decision theory and some useful incentive properties.

When the goodnesses come from the quadratic scoring rule, the different choice models imply distinctive predictions for continuous choice. The CR choice pattern is diagonal when the evidence stage is noiseless and increasingly undershoots as noise at the evidence stage increases. The RGR choice pattern is close to the uniform distribution even when the evidence stage is noiseless and is relatively insensitive to noise at the evidence stage.[1] Logit choice always undershoots in the continuous response mode, and the degree of undershooting increases with noise at the evidence stage or the decision stage. See the Appendix for formal derivations of these statements.

Table 1 summarizes the theoretical predictions. All three choice models predict that increased noise at the evidence stage will decrease overshooting or increase undershooting. In the limit, as noise completely dominates, the signal y is virtually independent of x, and we get extreme undershooting, as in the uniform distribution. The choice models differ in their basic patterns and in their predictions regarding noise at the decision stage. For example, if we find no impact for treatments designed to decrease decision stage noise and if binary choice patterns are approximately diagonal and continuous choice patterns nearly constant (uniform), we would conclude that the plain RGR is a good model of the decision process.

METHOD

We tested our predictions using variants of the well-known medical diagnosis task. As explained below, the main novelties are that we used two symptoms each with four levels rather than three or more binary-valued symptoms, as in most previous work; we used an expanded factorial design for the symptom configurations rather than a factorial design with interspersed trial blocks of single symptoms; and we introduced new treatments intended to manipulate noise levels at the evidence stage and at the decision stage.

Subjects. A total of 123 undergraduates from the University of California at Santa Cruz participated in this experiment in fulfillment of a class requirement. The subjects, who were enrolled in either of two lower division psychology classes, signed themselves up for this specific experiment. One third of these subjects also received pay for their participation: two of the six treatment cells in the experiment involved the distribution of pay on the basis of individual performance. Individual earnings during the experiment ranged from $5 to $17, with a mean of $13.63. The entire experimental procedure was approximately 2 h in length.

Table 1
Theoretical Predictions of the Three Choice Models

Decision Rule	Response Mode	Basic Pattern	Evidence Noise	Decision Noise
1. MR/CR	binary	step	overshoot less	NA
MR/CR	continuous	diagonal	undershoot more	NA
2. Plain RGR	binary	diagonal	undershoot more	NA
Plain RGR	continuous	near constant	negligible	NA
3. Logit	binary	overshoot (for $\lambda > 2$)	overshoot less	overshoot less
Logit	continuous	undershoot	undershoot more	undershoot more

Note—Basic patterns for scatterplots (of mean actual response vs. objective values) are predicted for each decision rule and response mode. *Step* refers to extreme overshooting, *constant* refers to extreme undershooting, and *diagonal* refers to no over- or undershooting. The impact of noise at the evidence and the decision stages is also predicted.

Apparatus. The experiment utilized a graphics computer program written in C++ and was conducted with Power Macintosh 7500/100 computers with full color monitors. Four sound-dampened isolated testing rooms were used, with 1 subject per room. The subjects viewed the experimental events on the monitor and responded via clicking the mouse on various icons in the display.

Procedure. The subjects were told that they would be diagnosing a series of fictitious patients on the basis of medical charts, which consisted of values on the symptoms of temperature and blood pressure (sometimes with one symptom missing). They were told that this was a learning experiment in which the goal was to learn the associations between symptoms and diseases. It was explained that the experiment reflected the real-life fact that the relationship between particular symptoms and a specific disease is sometimes weak or strong but never entirely certain. The subjects were informed that the experiment was designed to be difficult (especially in the beginning), that they could expect to make mistakes, and that the presentation order of the diseases was random. The subjects were asked not to take notes during the experiment, and this was monitored.

Stimuli. Each trial began with the presentation of the symptom values on the left side of the monitor display. The symptom values were displayed using two thermometer icons labeled temperature and blood pressure (see Figure 4, top panel). An enclosed area of each thermometer was filled with red to indicate the particular level of symptom that was present. Each symptom could take on one of four different values: high, medium high, medium low, and low. For example, a ¼ filled thermometer represented a low temperature or blood pressure, and a ¾ filled thermometer represented a medium-high temperature or blood pressure. Also present on the monitor were icons representing the two possible diseases: autochus and burlosis. The previous cases information could be accessed by the subject at this point (if available). The subjects selected a choice as to which disease he/she believed that the current patient possessed via clicking the mouse on the appropriate icon. A second response was then collected in the form of a confidence rating. The subjects were able to rate their confidence in their diagnosis for each patient by clicking the mouse along a response continuum, represented by a vertical bar, with the endpoints labeled *Not at All Confident* and *Completely Confident*. After both responses, an updated display was presented that indicated by a blue outline of its icon which of the two diseases the patient did in fact have. Also presented was the scoresheet (in the four conditions in which score was included). After viewing the scoresheet (if present), the subjects would move to the next trial via another mouse click.

Expanded factorial design. Table 2A–2B present the expanded design used in our experiment. As in a single-factor design, each of the symptoms is presented unimodally, for a total of 4 + 4 = 8 symptoms. As in the factorial design, each of the four temperature symptoms is combined with each of the four blood pressure symptoms for another 16 symptom configurations. Thus, there are a total of 24 configurations. The entries in Table 2A show the number of trials with each disease for each symptom configuration.

The experimental session consisted of a total of 480 trials. The number of observations for the different symptom configurations ranged from 11 to 33, with 18 of the 24 being between 16 and 26 observations. These frequencies were determined by the chosen symptom likelihoods shown in Table 2B. The same randomized presentation order of stimuli was used for all the subjects. The experiment was subject paced. The subjects were told this fact and also that all previous subjects had finished in less than the 2 allotted h. (This information helped limit the subjects' tendency to hurry through the experiment, especially toward the end.) The experiment was broken into three blocks of 160 trials, and the subjects were permitted 5-min breaks between blocks.

Treatments. A 2 × 3 factorial between-subjects design was used, with 20 subjects in each cell. (Attendance fluctuations gave us 2 extra subjects in one cell and 1 extra in another.) The evidence treatment had two conditions, history and no history, and the decision treatment had three conditions, pay + score, score, and no score, as explained below.

The history condition is intended to reduce noise at the evidence stage. It gave the subjects the option on each trial, before making their response, to view a chart of relevant cases previously encountered. If selected by clicking a *previous cases* icon, the chart stayed on the screen until the subject finished viewing it and clicked an *OK* icon. The chart displayed the number of previous patients with each disease that had possessed the symptom levels present in the current patient, as in Figure 4 (middle panel). The subjects in the no history condition had no access to such a chart.

The score condition is intended to reduce noise at the decision stage. It involved the computation of a score calculated in each trial from the continuous response $r \in [0,1]$ and the actual disease $d = 1$ (autochus) or $d = 0$ (burlosis), using the quadratic scoring rule $S(r,d) = A - B(r - d)^2$, with $A = 80$ and $B = 280$. The maximum possible score on a trial (correct binary response with complete confidence, so $r = d = 0$ or 1) was $A = 80$ points. The lowest possible score (incorrect binary response, with complete confidence so $|r - d| = 1$) was $A - B$ = negative 200 points. A *not at all* confident answer, coded $r = .5$, always resulted in $A - .25B = 10$ points. See the Appendix for more discussion of the quadratic scoring rule.

In the score and pay + score conditions, following each trial the screen presented the subject's score on that trial and the cumulative score to that point of the experiment, as in Figure 4 (bottom panel). Each subject in the pay + score condition was also paid $1 per 1,000 points of his or her final cumulative score at the end of the experiment. The payment procedures in this condition were explained immediately before to the beginning of the experiment. In the no score condition, no scores were presented, and no pay was given or discussed.[2]

RESULTS

In a given set of trials with symptom configuration s, let $P(s)$ be the mean binary response (the fraction of tri-

Table 2A
Expanded Factorial Design:
Instances of Disease (A, B) by Symptom Configuration

	Blood Pressure				
Temperature	High	Medium High	Medium Low	Low	None
High	3,8	10,5	19,2	32,1	16,4
Medium high	2,16	7,10	14,5	24,2	12,8
Medium low	2,24	5,14	10,7	16,2	8,12
Low	1,32	2,19	5,10	8,3	4,16
None	2,20	6,12	12,6	20,2	

Note—The first entry in the top row, for example, means that there were three cases of autochus (disease A) and eight cases of burlosis (disease B) in medical charts showing high temperature and high blood pressure; the last entry in the top row indicates 20 charts (16A and 4B) with no blood pressure reported and high temperature.

als in which $d = 1$, or autochus is chosen) and let $R(s)$ be the mean continuous response. As usual, let $x(s)$ represent the true (Bayesian posterior) probability of autochus, given s. We group each subject's responses into Block 1 = Trials 1–240 and Block 2 = Trials 241–480. For each of the 24 symptom configurations s, we tabulate $R(s)$ and $P(s)$ separately for each of the 123 subjects (in the 2×3 different conditions) in each of the two blocks. Thus, we have $24 \times 123 \times 2 = 5{,}904$ summary observations of $P(s)$ and another 5,904 observations of $R(s)$.

The scatterplot in Figure 5 illustrates the summary data for the subject with the highest earnings in the history with pay + score condition. Note that this subject tended to overshoot in binary choice and to undershoot in continuous choice. Figures 6a–6f show aggregate scatterplots, in which $R(s)$ and $P(s)$ are averaged across all subjects (and both blocks) in a given condition. The fitted curves will be explained shortly. Again, we have undershooting in continuous choice and overshooting in binary choice. The reader may be able to detect regularities consistent with the main theoretical predictions in these scatterplots, but a more systematic analysis is in order.

We first construct a direct summary measure of undershooting or overshooting the Bayesian posteriori probabilities. We define the sign indicator $i(s) = sgn(x(s) - .5)$, so i is +1 if autochus is more likely and −1 if burlosis is more likely, given s. The measure is the signed deviation $D_B(s) = (x(s) - P(s))i(s)$ for binary choice and $D_C(s) = (x(s) - R(s))i(s)$ for continuous choice. Notice that D is zero for probability matching and is positive for overshooting with mean value (over x in [0,1]) of .25 in the case of extreme overshooting—that is, for the step function at .5. Likewise, D is negative for undershooting and has a mean value of −.25 in the case of extreme undershooting—that is, for the uniform or constant response pattern.

Figure 7 reports the mean values of D, averaged over the 24 symptom conditions, as a function of block, evidence condition, decision condition, and response mode. According to the present theoretical framework, we expect main effects for each of these variables, as well as interactions between evidence condition and response mode and between decision condition and response mode. To provide a statistical evaluation of these effects, a $2 \times 2 \times 3 \times 2 \times$ ~20 analysis of variance (ANOVA) was carried out on this measure, with block, evidence condition, decision condition, response mode, and subjects as factors. The mean values vary considerably and cover much of the possible range between −0.25 and +0.25. The most striking result is the effect of response mode [$F(1,117) = 1{,}385, p < .001$]: all entries are positive for binary choice, implying overshooting, whereas most entries are negative for continuous choice, implying undershooting. This is consistent with the Logit choice model.

Every Block 2 entry is larger than the corresponding Block 1 entry [$F(1,117) = 311, p < .001$], suggesting that underconfidence and probability matching diminish with experience. Every history bar is higher than the corresponding no history bar [$F(1,117) = 362, p < .001$], consistent with theoretical predictions. Finally, every score bar is higher than the corresponding no score bar [$F(1,117) = 89, p < .001$], consistent with the Logit model. However, the pay + score bars bear no consistent relation to the corresponding score bars; in a few cases they are higher, but more often they are lower or about the same. The associated ANOVA results [$F(1,117) = 49, p < .001$] indicate that pay + score produces, on balance, less overshooting than does score, but the overall main effect of the score treatment was positive and significant [$F(2,117) = 47.5$, $p < .001$].

Only two interactions were statistically significant ($p < .01$): the two-way interaction of decision condition

Table 2B
Expanded Factorial Design: Objective Symptom Likelihoods

	p(Temperature|Disease)			p(Blood Pressure|Disease)		
Symptom Level	A	B	Log Odds	A	B	Log Odds
High	.4	.1	1.39	.05	.5	2.40
Medium high	.3	.2	0.41	.15	.3	0.69
Medium low	.2	.3	−0.41	.3	.15	−0.69
Low	.1	.4	−1.39	.5	.05	−2.40

Note—The underlying likelihoods used to generate the medical charts. For example, when a chart for a patient with disease A shows temperature, it reports a high level with probability .4. With disease B, the corresponding probability is .1, and the log-odds are ln(.4/.1) = 1.39.

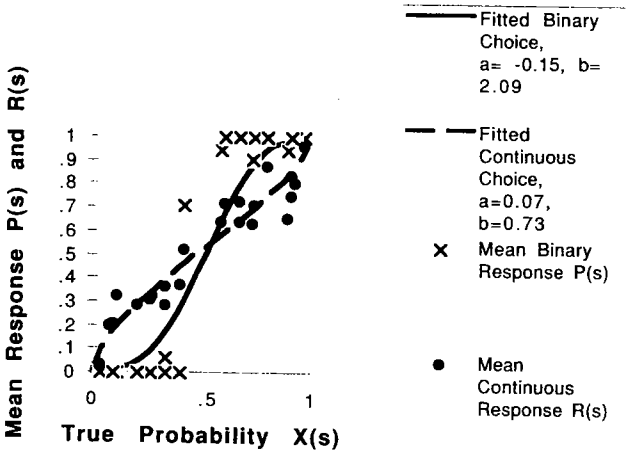


Figure 5. Scatterplot of responses $P(s)$ and $R(s)$ for Subject 28, who had the highest earnings in the history with pay + score condition. The fitted curves are explained later in the text.

and response mode [$F(2,117) = 9.39, p < .001$] and the three-way interaction of evidence, decision condition, and response mode [$F(2,117) = 7.89, p < .001$].

The results based on our measure of overshooting seem quite conclusive, but one potential difficulty comes to mind. The definition of the dependent variable implicitly assumes that all the subjects use the same unbiased critical value $c = .5$. If a subject were biased to respond with one of the two diseases, the D value would not properly measure under- or overshooting for that subject.

To resolve the difficulty, we constructed for each subject and decision mode a more sophisticated measure of overshooting. It is based on the log odds transformation[3] $L(q) = \ln q - \ln(1 - q)$, the inverse of the unit logistic transformation $S(x|\lambda = 1) = (1 + \exp(-x))^{-1}$. The idea is that transforming the data $P(s)$ and $R(s)$ and the explanatory variable $x(s)$ by L permits us to estimate plausible linear relationships that, transformed back by $L^{-1} = S$, display clearly the degree of over- or undershooting. The measure generalizes that of McDowell and Oden (1995) by allowing the critical value c to vary freely.

For each subject and each block, we run the regression

$$L(P(s)) = a + bL(x(s)) + e(s) \qquad (2)$$

on the 24 data points given by the different values of s. The regression gives us coefficient estimates $\hat{a}$ and $\hat{b}$ for binary choice for that subject and block. We also run the corresponding continuous regression, with dependent variable $L(R(s))$. The fitted curve $\hat{P}(x) = S(\hat{a} + \hat{b}L(x))$ displays directly the degree of overshooting in binary choice, and the corresponding curve $\hat{R}(x)$, using the coefficient estimates from the $L(P(s))$ regression, shows the degree of overshooting in continuous choice. The fitted curves are illustrated in Figures 6a–6h.

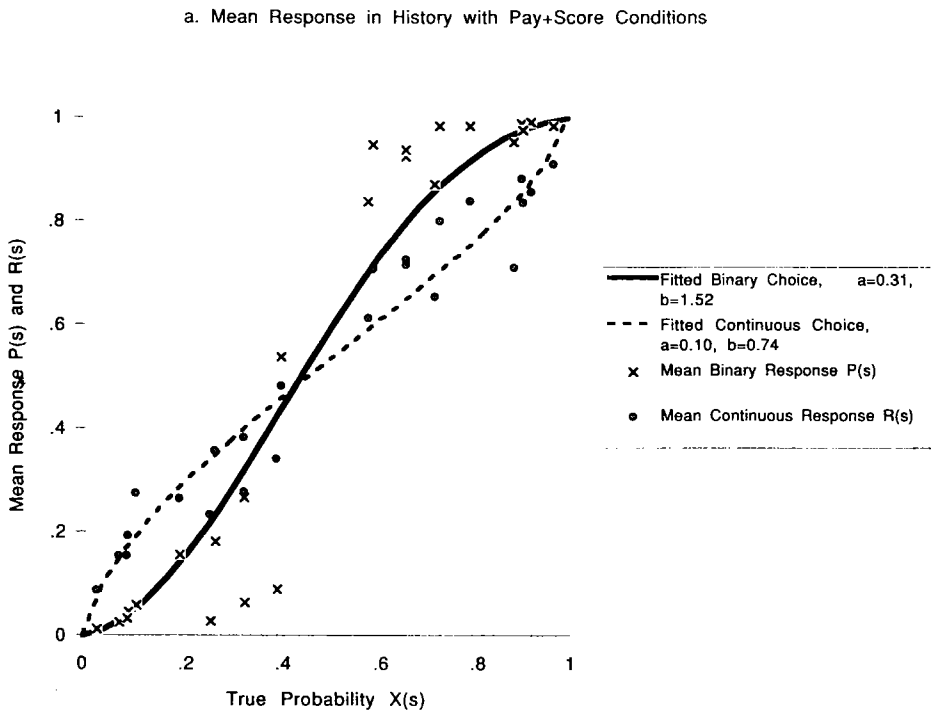
One can infer the degree of overshooting and the critical value from the estimated coefficients $\hat{a}$ and $\hat{b}$. Since $S(\hat{a})$ is the height of the fitted curve $\hat{P}(x)$ or $\hat{R}(x)$ at $x = .5$, we see that $\hat{a}$ has inverse relation to c. Positive values of $\hat{a}$ imply a bias toward $r = 1$, and negative values imply a bias towards $r = 0$. The coefficient estimate $\hat{b} = 0$ implies that $L(P)$ is constant; hence, P is constant, and we have extreme undershooting. The estimate $\hat{b} = 1$ implies that $L(P)$ increases 1 for 1 with $L(x)$; hence, (ignoring any bias from nonzero $\hat{a}$) the choice pattern is diagonal. For very large values of $\hat{b}$, the fitted curve $\hat{P}(x)$ is close to the unit step function. In short, $\hat{b} > 1$ implies overshooting, and $\hat{b} < 1$ implies undershooting.

Figure 8 presents the mean estimates $\hat{b}$. The patterns of overshooting and undershooting are very similar to those of the cruder measure D reported in Figure 7. The mean estimates for binary choice range from 1.25 to 2.0, implying clear overshooting, whereas those for continuous choice are mostly below 1.0 and all below 1.25, implying undershooting or approximate probability-matching behavior. Indeed, mean behavior is almost precise probability matching ($\hat{\beta} = .99$) in the least noisy continuous response case, Block 2 with history and score. Estimates always increase as we move from Block 1 to Block 2, from no history to history, or from no score to score. In agreement with the ANOVA on the overshooting measure D, all main effects and the relevant interactions were statistically significant ($p < .001$).

Other Evidence

Besides data in the $2 \times 3 = 6$ conditions just described, we have pilot experiment data from 26 subjects. The stimuli and response alternatives were exactly the same but presented in a slightly different format. The subjects in this seventh cell had no history and no score. We also analyzed binary and continuous choice data aggregated over the 60 subjects in the medical diagnosis experiment of Nosofsky et al. (1992), as listed in their Table 2. Again, the treatments were no history and no score.

Figures 6g–6h plot the two supplementary data sets and corroborate our findings from the main data set. The Nosofsky et al. (1992) experiment a priori should be nois-



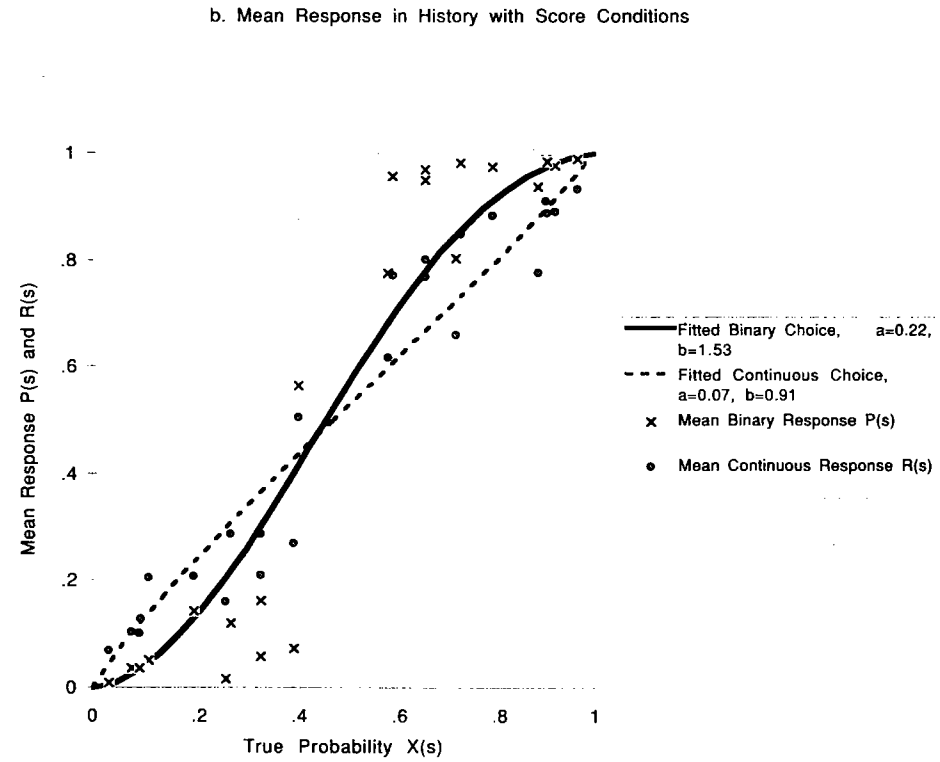
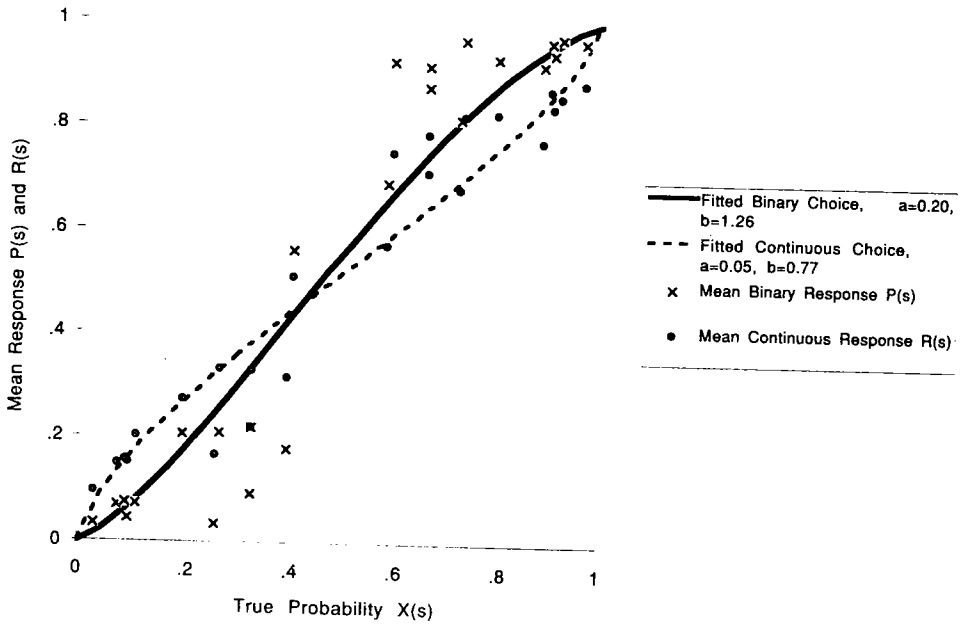


Figure 6. Scatterplots of subjects' responses *R(s)* and *P(s)* averaged across all subjects (and both blocks) in a given condition. The fitted curves are explained in the text. Panels a–f give the data for the present experiment. Panel g is from "Combining Exemplar-Based Category Representations and Connectionist Learning Rules," by R. M. Nosofsky, J. K. Kruschke, and S. C. McKinley, 1992, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18,* 211-233 (Table 2); copyright 1992 by the American Psychological Association. Adapted with permission. Panel h is from a pilot experiment.

c. Mean Response in History with No Score Conditions



d. Mean Response in No History with Pay+Score Conditions

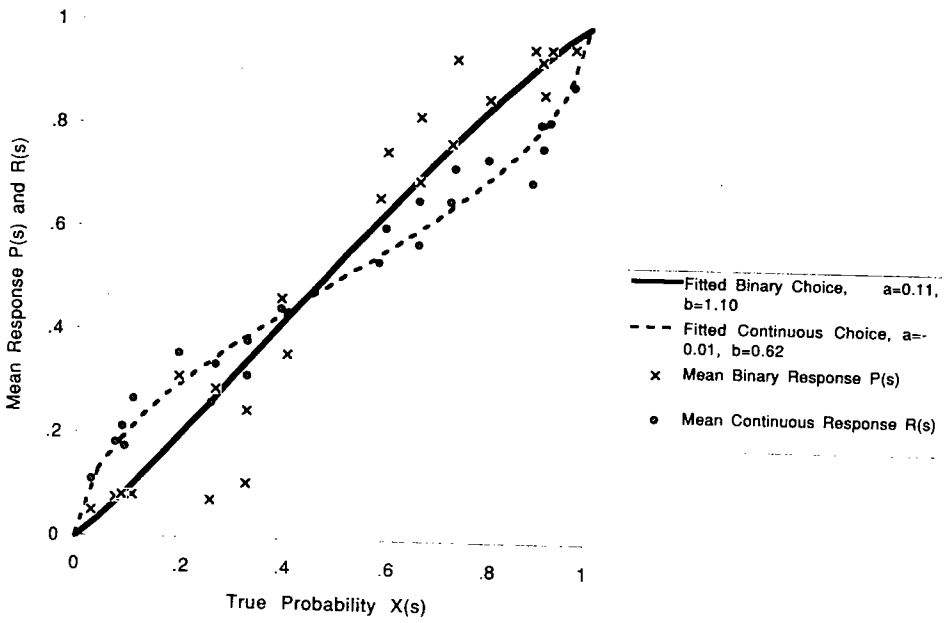
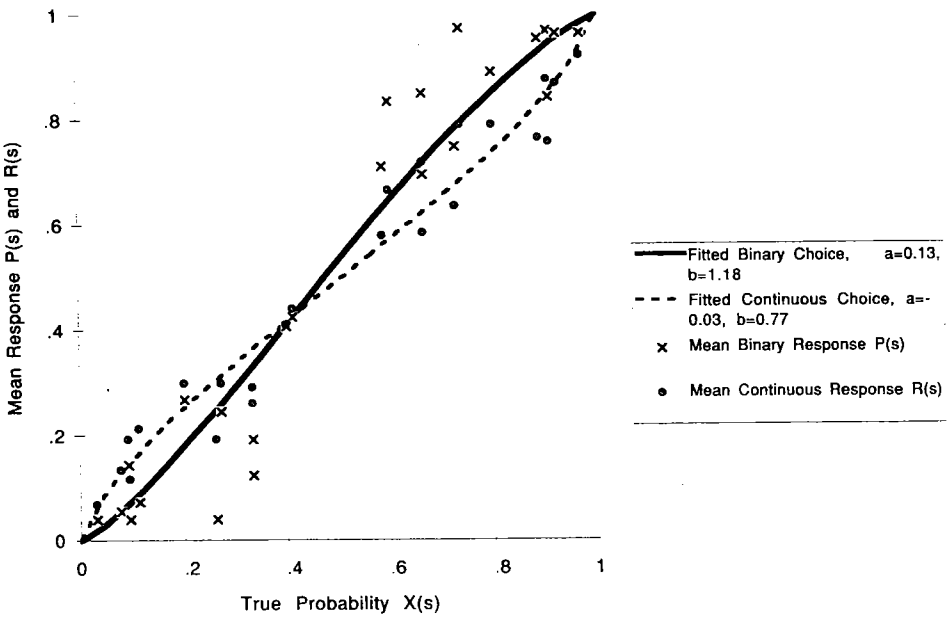


Figure 6 (continued).

e. Mean Response in No History with Score Conditions



f. Mean Response in No History with No Score Conditions

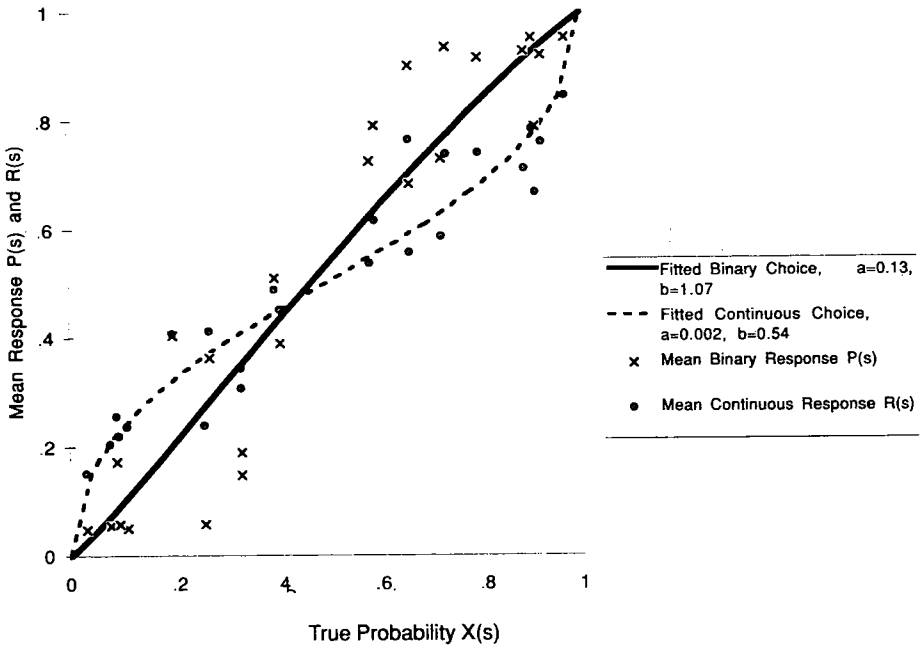
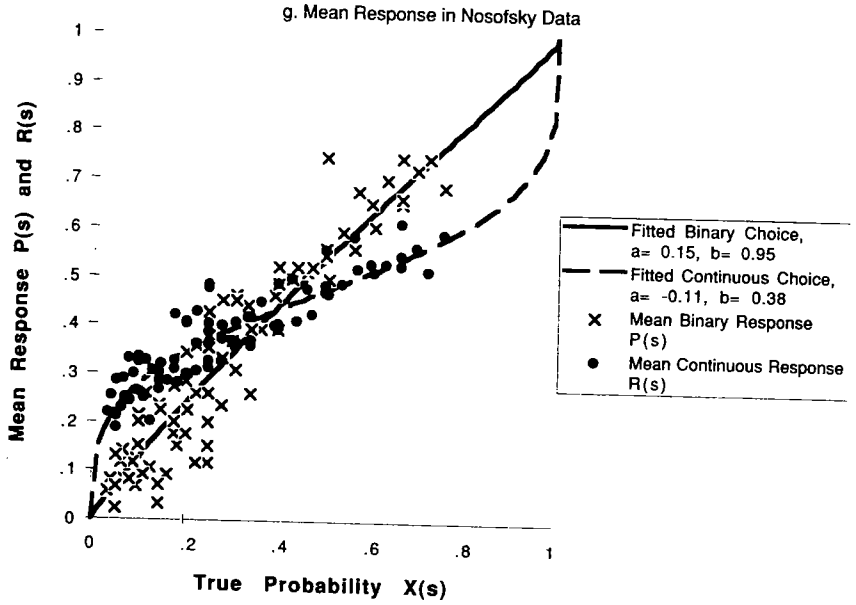


Figure 6 (continued).



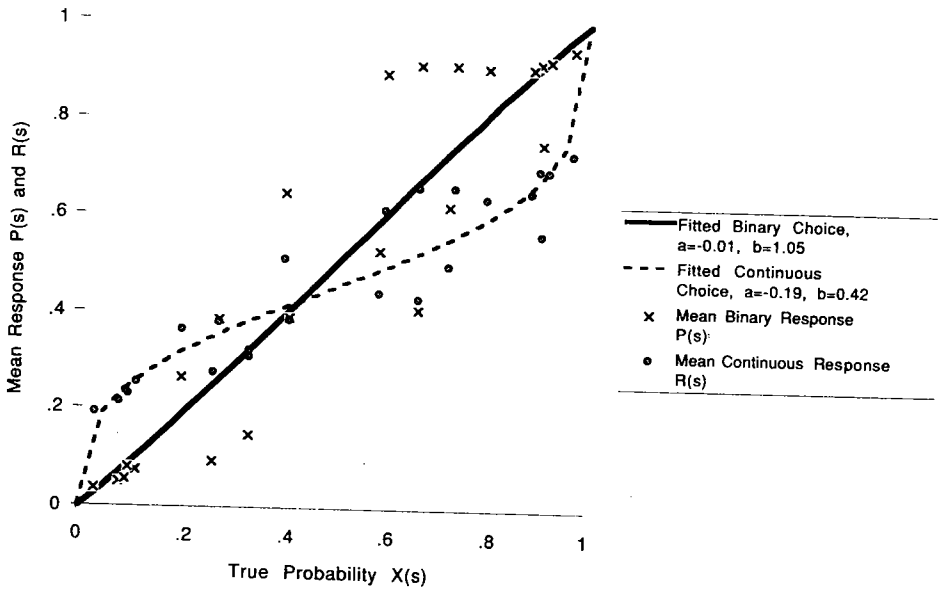


Figure 6 (continued).

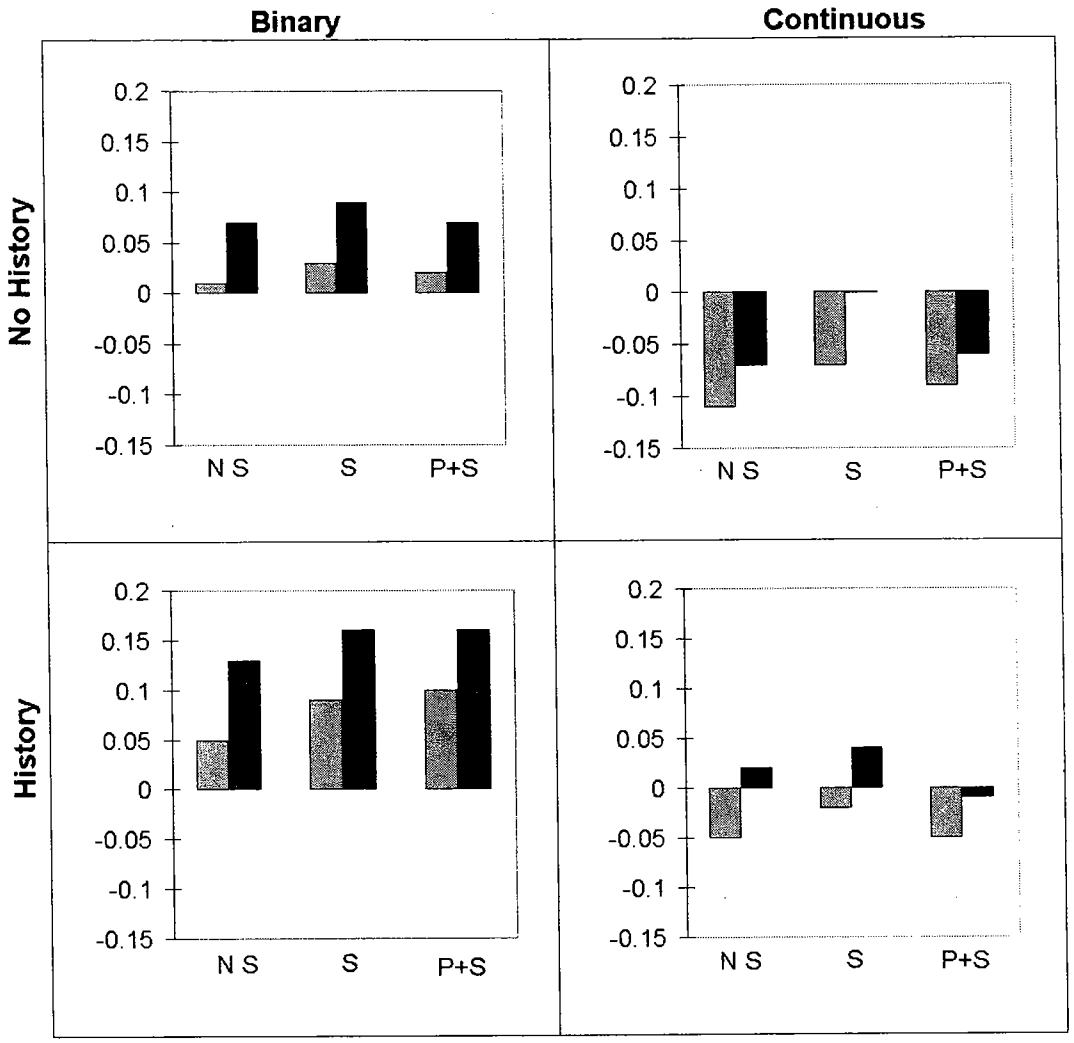


Figure 7. The mean values of the overshooting measure *D* averaged over the 24 symptom conditions, as a function of evidence condition (history vs. no history), decision condition (NS = no score vs. S = score only vs. P + S = pay + score), and response mode. The light bars correspond to Block 1, and the dark bars, to Block 2.

ier than ours, because they used a more complicated set of stimuli (81 configurations with unequal prior probabilities) and no noise reduction treatments. The $\hat{b}$ slope estimates from their data are indeed smaller than those in our data: the fitted continuous curve undershoots greatly ($\hat{b} = .384$), and the fitted binary curve is almost diagonal ($\hat{b} = .948$). Our pilot data also are a priori quite noisy, and the results are quite similar, with $\hat{b} = .374$ for continuous and .932 for binary choice.

A potential problem with the analysis in the previous subsection is that the experimental design manipulated the decision and evidence conditions as between-subjects independent variables. Individual subject variability, thus, could have washed out the influence of evidence and decision treatments, or, on the other hand, the estimated treatment effects could have been due to a few aberrant individual subjects. To resolve the problem with the broadest possible assessment, we combine Block 1 and Block 2 responses and also look at the pilot data. We screen out a few of the most erratic subjects—those whose choices were not systematic, as indicated by an R^2 for regression Equation 2 below .50. Table 3 counts the number of remaining subjects whose coefficient estimates fall into the relevant ranges. The ranges corresponding to points $\hat{a} = 0$ and $\hat{b} = 1$ include one standard error on either side of the point. Almost a fifth of our systematic subjects show a bias toward $r = 1$ (29 of 138 in binary choice and 26 of 142 in continuous choice), and a few show a bias toward $r = 0$, but a clear majority fall into the unbiased category $\hat{a} = 0$.

Table 3 generally confirms undershooting in continuous choice and overshooting in binary choice. In binary choice, 115 subjects overshoot ($\hat{b} > 1$), whereas only 19 probability match and only 4 undershoot, despite the rather generous convention for probability matching ($b = 1$) that includes a standard error either way; typically $\hat{b}$ between 0.8 to 1.2 is classified as $b = 1$. In continuous

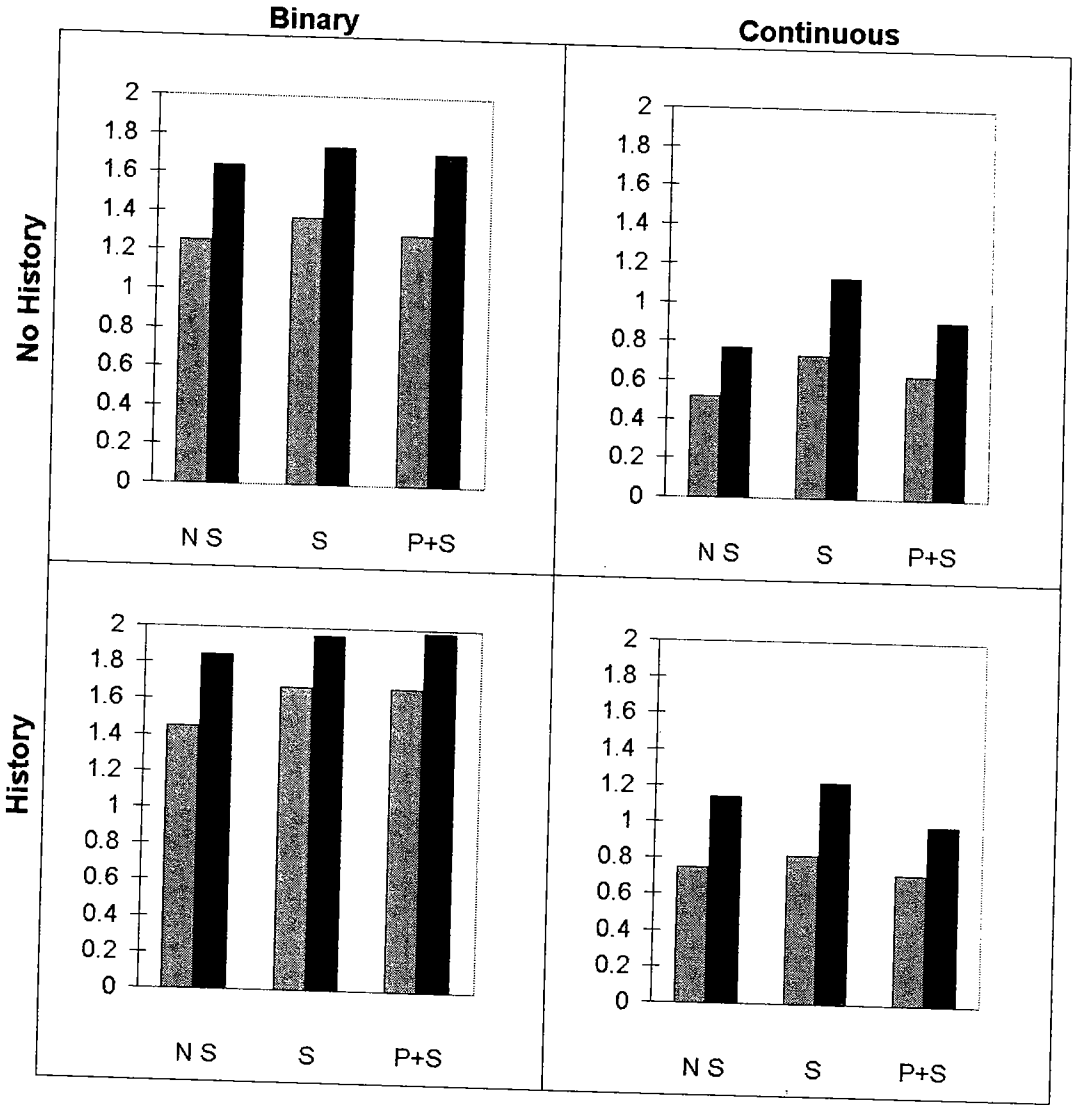


Figure 8. The mean values of overshooting measure $\hat{b}$. Description same as in Figure 7.

choice, undershooting ($\hat{b} < 1$) dominates, with 91 such subjects versus 26 probability matchers and 25 overshooters. Combining the evidence, we get a very strong result from panel C of the table: 147 of the subjects overshoot more in binary than in continuous tasks, none are within a standard error of overshooting equally or less, and only 2 subjects had ill-fitting regressions.

In short, all the evidence we have seen corroborates the inferences drawn from Figure 7: As predicted by the Logit model, continuous choice tends to undershoot, especially in noisier conditions, and binary choice tends to overshoot, especially in less noisy conditions.

Our data bear on another related question. Are continuous choice data reliable? We have already noted the reservations of Nosofsky et al. (1992), which are widespread among psychologists. It is true that only in the least noisy conditions do we find continuous choice on the diagonal, where it can be regarded as a direct report of subjective probability. However, a simple transformation on the basis of Equation 2 brings the continuous data into close alignment with objective probabilities. For 113 of 123 test subjects (and for 23 of 26 pilot subjects), the R^2 is higher for the continuous choice data than for the binary choice data in Equation 2 regressions; the binomial (or signs) test indicates a significant difference ($p < .001$). We conclude that, in a meaningful sense, the continuous data are more internally consistent than the binary choice data.

DISCUSSION

Our analysis points to several conclusions. First, the data clearly exhibit undershooting in continuous choice (or ratings). The degree of undershooting is systematically reduced by treatments designed to reduce noise at the evidence stage and at the decision stage. Indeed, undershooting in continuous choice virtually disappears in the