Psychonomic Bulletin & Review
2008, 15 (3), 465-494
doi: 10.3758/PBR.15.3.465

THEORETICAL AND REVIEW ARTICLES

Decision noise: An explanation for observed violations of signal detection theory

SHANE T. MUELLER
Indiana University, Bloomington, Indiana

AND

CHRISTOPH T. WEIDEMANN
University of Pennsylvania, Philadelphia, Pennsylvania

In signal detection theory (SDT), responses are governed by perceptual noise and a flexible decision criterion. Recent criticisms of SDT (see, e.g., Balakrishnan, 1999) have identified violations of its assumptions, and researchers have suggested that SDT fundamentally misrepresents perceptual and decision processes. We hypothesize that, instead, these violations of SDT stem from decision noise: the inability to use deterministic response criteria. In order to investigate this hypothesis, we present a simple extension of SDT—the decision noise model—with which we demonstrate that shifts in a decision criterion can be masked by decision noise. In addition, we propose a new statistic that can help identify whether the violations of SDT stem from perceptual or from decision processes. The results of a stimulus classification experiment—together with model fits to past experiments—show that decision noise substantially affects performance. These findings suggest that decision noise is important across a wide range of tasks and needs to be better understood in order to accurately measure perceptual processes.

Signal detection theory (SDT) has become a prominent and useful tool for analyzing performance across a wide spectrum of psychological tasks, from single-cell recordings and perceptual discrimination to high-level categorization, medical decision making, and memory tasks. The utility of SDT comes from its clear and simple account of how detection or classification performance can be translated into psychological quantities, such as sensitivity and bias. Whether its use is appropriate for a specific application depends on a number of underlying assumptions, and even though these assumptions are rarely tested, SDT has proved useful enough that it is considered one of the great successes of cognitive psychology. Yet, SDT has also undergone criticism, which began to emerge when this theory was relatively young.

Criticisms of SDT

SDT assumes that percepts are noisy and give rise to overlapping perceptual distributions for signal and noise trials. In order to distinguish between signal and noise trials, the observer uses a decision criterion to classify the percepts. Signal responses are "hits" when they are correct and "false alarms" when they are incorrect; similarly, noise responses can be classified as "correct rejections" and "misses." Many criticisms of SDT have centered on how the observer places a decision criterion during a detection or classification task, and whether a deterministic criterion is used at all (see, e.g., Dorfman & Biderman, 1971; Dorfman, Saslow, & Simpson, 1975; Kac, 1969; Kubovy & Healy, 1977; Larkin, 1971).

Clearly, when initially performing a signal detection task,[1] an observer may be unable to estimate stimulus distributions and payoff values accurately; thus, one might expect the placement of a decision criterion to improve with experience, approaching a static optimal criterion. Yet, some results suggest that even with extensive practice, responses can be suboptimal: There are numerous demonstrations of human probability micromatching in signal detection tasks (see, e.g., Dusoir, 1974; Lee, 1963; Thomas, 1973, 1975) and other demonstrations that static decision criteria are not typically used (e.g., Healy & Kubovy, 1981; Lee & Janke, 1964; Lee & Zentall, 1966; Treisman & Williams, 1984). Despite the fact that models accounting for these dynamics are based on a fairly reasonable assumption (i.e., that the decision criterion should improve with experience), they have not enjoyed the success of classic SDT—probably because they add layers of complexity to the theory that are not easily accommodated or validated. Given that even the basic assumptions required by SDT are rarely tested, it is perhaps not surprising that tests of these additional factors happen even less frequently.

S. T. Mueller, smueller@ara.com

More recently, Balakrishnan (1998a, 1998b, 1999) raised new objections to SDT on the basis of consistent violations of its assumptions: (1) Receiver operating characteristic (ROC; see below) functions produced under different base rates have different shapes (whereas SDT predicts that they should lie on top of one another), and (2) confidence-based measures typically indicate no change of the decision criterion in response to base rate manipulations (see, e.g., Balakrishnan, 1999). Balakrishnan's criticisms differ from the earlier criticisms discussed previously, because he did not simply suggest that the violations of SDT are due to a suboptimal criterion placement or similar imperfections within the framework of SDT. Instead, he claimed that they expose fundamental flaws in the most basic assumptions of SDT. Therefore, his criticism calls into question the results from thousands of published studies that have relied on SDT's assumptions to quantify perceptual and decision processes.

In this article, we will examine the violations of SDT and argue that they could stem from *decision noise*—uncertainty in the mapping between an internal perceptual state and the overt response. Furthermore, we will present a new extension of SDT—the *decision noise model* (DNM)—that incorporates decision noise and perceptual factors in signal detection tasks. We will also introduce a new type of ROC function that can be used in conjunction with conventional confidence-based ROC functions to distinguish perceptual and decision processes. Our application of this ROC analysis to the data of a new stimulus classification experiment—along with the fits of the DNM to these data and those collected by Balakrishnan (1998a) and Van Zandt (2000)—suggest that decision noise needs to be acknowledged as a primary source of noise in signal detection.

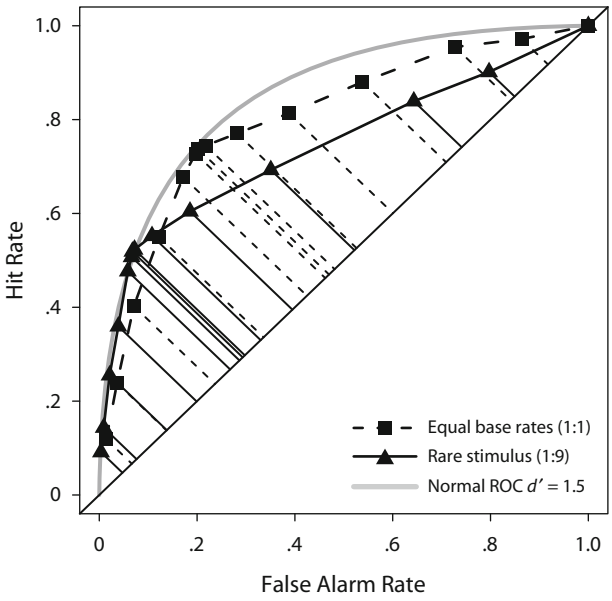

Figure 1. Two confidence ROC (C-ROC) functions based on data from the mixed successive condition of Balakrishnan (1998a) for equally likely signal and noise trials versus rare signal trials. The points on each C-ROC show the hit and false alarm rates for each of the 14 confidence levels (7 for each response). The functions cross one another, apparently violating the assumption of SDT that such manipulations affect only the decision criterion. For comparison, a normal ROC function with a d' of 1.5 is plotted as well. The length of the lines connecting ROC points to the $y = x$ diagonal are proportional to values of the $U_R(k)$ functions for each condition.

Confidence ROC Functions Can Change Shape

According to classic SDT, the perceptual distributions of the signal and noise trials form a regime under which a decision rule operates. These distributions are determined by the stimulus and the perceptual system, but are otherwise relatively fixed. In contrast, the observer has strategic control over the decision criterion, which may be placed at an optimal position in order to maximize accuracy or payoff. A single observer might adopt a very strict criterion in some situations, producing very few false alarms but also few hits; in other situations, the criterion may be lax, producing many hits but also many false alarms. Because the criterion is under strategic control, the observer might use a suboptimal strategy, either moving the criterion to an inappropriate location, or even placing the criterion at a different location on each trial. Standard SDT statistics can easily deal with nonoptimal placement of a static criterion, but a nonstatic response criterion introduces noise that is attributed to perceptual rather than to decision processes. This represents a potential weakness of the model, or at least the derived statistics d' (sensitivity) and β (bias), which assume a fixed decision criterion.

ROC functions are sometimes measured to verify whether the assumptions of SDT are valid. To do this, the experimenter manipulates instructions, base rates, or payoffs, in order to encourage the observer to adopt different response policies. The ROC function is formed by plotting hit rate against false alarm rate for these different conditions. The form of this function can be compared with theoretical functions generated from Gaussian distributions to determine whether the distributional assumptions of the model are appropriate. Estimating an ROC function this way is costly and time consuming; thus, a more efficient procedure relying on confidence ratings is often used. Rather than asking the observer to adopt a single confidence criterion throughout a task, one instead asks for a confidence rating qualifying each response, which can then be used as a stand-in for different criterion levels. The resulting confidence ROC (C-ROC) function also enables other factors, such as base rate, to be manipulated simultaneously, thereby allowing another fundamental assumption of SDT to be tested: Manipulations of signal base rate or of response payoff should not change the perceptual distributions of signal or noise trials and should therefore produce C-ROC functions that lie on top of one another (although the points associated with specific confidence values may lie at different positions on the function).

Balakrishnan (1998a) conducted experiments that tested this prediction. He formed C-ROC functions for two conditions of a detection experiment: one in which the stimulus and noise trials were equally likely, and one in which the stimulus appeared on only 10% of the trials. As is shown

in Figure 1, the C-ROC functions he obtained differed substantially, apparently violating the assumption of SDT that manipulations of base rate affect only the decision criterion and not the shape of the perceptual distributions. Similar results have been obtained in other conditions and by other experimenters (see, e.g., Van Zandt, 2000).

We hypothesize that this violation of SDT may stem from the confidence rating procedure itself. Examining Figure 1, we find that not only do the C-ROC functions cross, but they each have a noticeable peak at or near the middle confidence point.[2] This point corresponds to the overall hit/false alarm rate for that condition if confidences were aggregated into just two responses using the middle confidence point. These central points can also be used to compute β for this experiment, and, for the rare signal condition, β moves in the direction expected if the observer placed a criterion in order to improve accuracy in response to a base rate manipulation. Furthermore, both peaks fall on the normal ROC function, with $d' = 1.5$. Consequently, if only the two-category classification responses were analyzed, this experiment would seem to support SDT: Under two different base rates, approximately equal sensitivity was observed, along with an appropriate change in decision criterion. The violations appear when we consider the confidence data. According to SDT, the C-ROC functions should have followed the same normal ROC contour in both conditions. This failure to conform to the assumptions of SDT may indicate that the underlying decision model is misspecified. However, it also may indicate that confidence ratings distort the evidence distribution and are therefore inappropriate for making conclusions about these perceptual distributions or decision processes that operate on them.

Noise in Signal Detection

As described previously, in SDT, it is assumed that stimuli give rise to noisy percepts. This uncertainty in the mapping of external stimuli to internal perceptual states is called *perceptual noise*, and it is the only source of noise considered in classical SDT. As perceptual noise increases, accuracy decreases. As accuracy decreases, the ROC function approaches the diagonal line $y = x$, which represents response states that do not discriminate signal from noise trials.

Noise might also be introduced in the mapping between the internal perceptual state and the response. Although this *decision noise* is typically not addressed in classic SDT, it clearly might exist and may have an impact on both binary stimulus classification and confidence assessment. In fact, peaked ROC functions—like the one shown in Figure 1—could occur if the noise in the mapping from perceptual evidence to confidence responses (henceforth called *confidence noise*) is relatively larger than that in the mapping from perceptual evidence to the binary response class (henceforth called *classification noise*). In the presence of decision noise, points on the C-ROC function are a mixture of multiple points on a latent perceptual ROC function and thus lie below the ROC function that would be formed if deterministic decision criteria were kept constant within an experimental condition.[3]

In fact, each C-ROC function in Figure 1 appears to be well approximated by a two-piece linear function from (0,0) to (1, 1) through the middle confidence point corresponding to the classification criterion. A bilinear function like this would be obtained if confidence ratings within each classification category were simply randomly assigned without regard to any internal level of evidence. Thus, different levels of classification and confidence noise entering into the decision process may produce the puzzling results that were noted by Balakrishnan (1998a, 1999), even if all other assumptions of SDT were correct. We will examine this possibility in greater detail below, but first we will turn to a related finding that also poses problems for SDT.

Lack of Evidence for Criterion Change

Balakrishnan (1998b) introduced several new statistics that allow better tests of criterion shifts of the type assumed by SDT. In this context, he proposed a function—$U_R(k)$—that measures the divergence between the cumulative density functions for the signal and noise distributions estimated at the transitions (k) between confidence responses. For each point on a C-ROC function, the associated $U_R(k)$ value is proportional to the distance between that point and the diagonal line $y = x$ along a vector with slope -1 (see Figure 1).[4] Because of this correspondence, $U_R(k)$ is closely related to the area under the ROC function, which is commonly used as an index of sensitivity. Balakrishnan (1998b) showed that if the decision criterion changes, the peak of the $U_R(k)$ function should move away from the central confidence point. If the C-ROC functions had followed the $d' = 1.5$ ROC contour but the confidence points had shifted along the contour, then the peak of the $U_R(k)$ function would have shifted as well. Despite the fact that the ROC functions differed between conditions, the peak of the $U_R(k)$ functions did not change; that is, it is always located at the central confidence point.

Balakrishnan and MacDonald (2002) noted that across a wide range of data, the peak of the $U_R(k)$ function rarely changed in response to manipulations that affected β. This result suggests that decision criteria do not actually shift position in the way assumed by SDT. Balakrishnan and MacDonald suggested that the decision criterion may remain fixed at an equal-likelihood point for the two distributions, whereas the variances (and/or shapes) of the signal and noise distributions may change in response to manipulations of payoff and base rate. For example, in a classic signal detection task, the variance of perceptual states produced during signal trials may be smaller than the variance of those produced during noise trials in a condition in which the signal occurs often. As noted by Balakrishnan and MacDonald, these types of changes in perceptual distributions are incompatible with SDT, but are natural consequences of a class of sequential sampling models.

Treisman (2002) noted several objections to these arguments, and Balakrishnan and MacDonald (2002) defended their utility. However, the arguments centered on the different ways in which a set of deterministic response criteria might interact to produce the observed results. In our assessment, the analyses by Balakrishnan (1998a, 1998b, 1999)

present substantial challenges for SDT and are not just complications caused by degenerate criterion placement, as was suggested by Treisman. However, we hypothesize that the apparent violations of SDT may stem from decision noise and, specifically, probabilistic response processes associated with confidence ratings. As we discussed previously, if the uncertainty involved in rating confidence (i.e., confidence noise) is relatively greater than the uncertainty in determining an overall classification category (i.e., classification noise), then the C-ROC function [and associated $U_R(k)$ functions] will be peaked at the point between the two classification categories. The central peak in the $U_R(k)$ function produced by this confidence noise could hide a shift in the function's peak that would otherwise result from a criterion shift. Thus, the apparent violations of SDT may not reflect fundamental misrepresentations of the classic SDT, but instead reflect inappropriate assumptions about how humans determine their confidence responses. In order to evaluate this possibility, we will next describe a new extension of SDT that incorporates decision noise and allows confidence noise and classification noise to vary independently.

THE DECISION NOISE MODEL (DNM)
A Signal Detection Model With Response Uncertainty

We hypothesize that Balakrishnan's (1998a, 1999) findings can be explained by noise entering into the decision process. In order to investigate this possibility, we have developed an extension of SDT that we call the DNM. This model incorporates both perceptual noise and decision noise, with independent contributions from classification and confidence noise. We use this model not as a replacement for SDT (and do not create new measures of sensitivity and bias based on it), but as an extension of classic SDT that can illustrate how different sources of noise may affect measurable statistics. This model incorporates confidence ratings and encapsulates aspects of decision uncertainty present in numerous previous models (see, e.g., Busemeyer & Myung, 1992; Erev, 1998; Kac, 1969; Schoeffler, 1965; Treisman & Williams, 1984), but does so at a level that does not incorporate learning and other trial-by-trial dynamics present in many of these previous models. This simplification allows us to evaluate the role of decision noise in general, independent of the specific assumptions of these theories (i.e., learning scheme, response mapping, criterion sampling/drift, etc.). We present an overview of the DNM next and a more detailed formal presentation in Appendix A.

Before we describe the model in greater detail, a discussion about one of its fundamental assumptions is necessary. Balakrishnan and MacDonald (2002) argued that the data we have described support a sequential sampling model. However, in this article, we will show that decision noise is also a reasonable explanation. Indeed, this is a false dichotomy: Reasonable models could be formed that produce an internal perceptual state using a sequential sampling process, but that still introduce decision noise in the mapping between this internal state and classification or confidence responses. Consequently, we are especially interested in developing both a model in which decision noise alone can account for the findings and a method to assess the role of decision noise in signal detection tasks with confidence ratings.

In our model, the nominal stimulus is a categorical variable describing the stimulus class. The actual distal stimulus presented to the observer may be a noisy exemplar from a stimulus class or a pure stimulus prototype presented in noise, so that even an ideal observer may be unable to attain perfect accuracy. When presented, the observer's percept may be a noisy function of the distal stimulus (as in SDT). Additionally, we will allow this function to differ for different stimulus classes in order to investigate the possibility that asymmetric C-ROC functions occur because of changes in the shape of the perceptual distributions. Finally, we assume that there is decision noise—a probabilistic mapping from percept onto response—so that even identical percepts may lead to different responses on different occasions. As described above, we distinguish between two components of decision noise: classification noise (noise in the assignment of a categorical response) and confidence noise (noise in the assignment of a confidence level).

Mapping From Distal Stimulus to Percept

We refer to internal noise that distorts the representation of the distal stimulus as perceptual noise and to the resulting distribution of perceptual states as the perceptual distribution. Traditionally, d' is attributed to the combined effect of external and perceptual noise, both of which affect the perceptual distributions. In addition, in SDT, it is typically assumed that signal and noise trials produce perceptual distributions with the same variance, and that these distributions do not change in response to base rate manipulations. However, Balakrishnan and MacDonald (2002) suggested that the observed crossover in ROC functions (see Figure 1) could stem from perceptual distributions that changed shape in response to manipulations of base rate or payoff. Consequently, we allow such changes to occur.

Mapping From Percept to Response

Decision noise is not consistent with a static decision criterion typically assumed in SDT, and the presence of decision noise would allow two identical internal percepts to produce different responses on different occasions. Decision noise has frequently been ignored because it often cannot be separated from perceptual noise and is simply incorporated into d', underestimating the level of perceptual sensitivity. There are many ways decision noise could be conceptualized (see Mueller, 1998, for a review). Some theorists have suggested that the decision criterion drifts along a sensory continuum from trial to trial, perhaps in response to error feedback (see, e.g., Kac, 1969). Others have suggested that decision criteria are sampled from a distribution on each trial (e.g., Erev, 1998), and still others have suggested that the observer learns a probabilistic function mapping sensory evidence onto the response (e.g., Schoeffler, 1965). Exactly how noise enters into the decision process is not important for our argument; thus, we as-

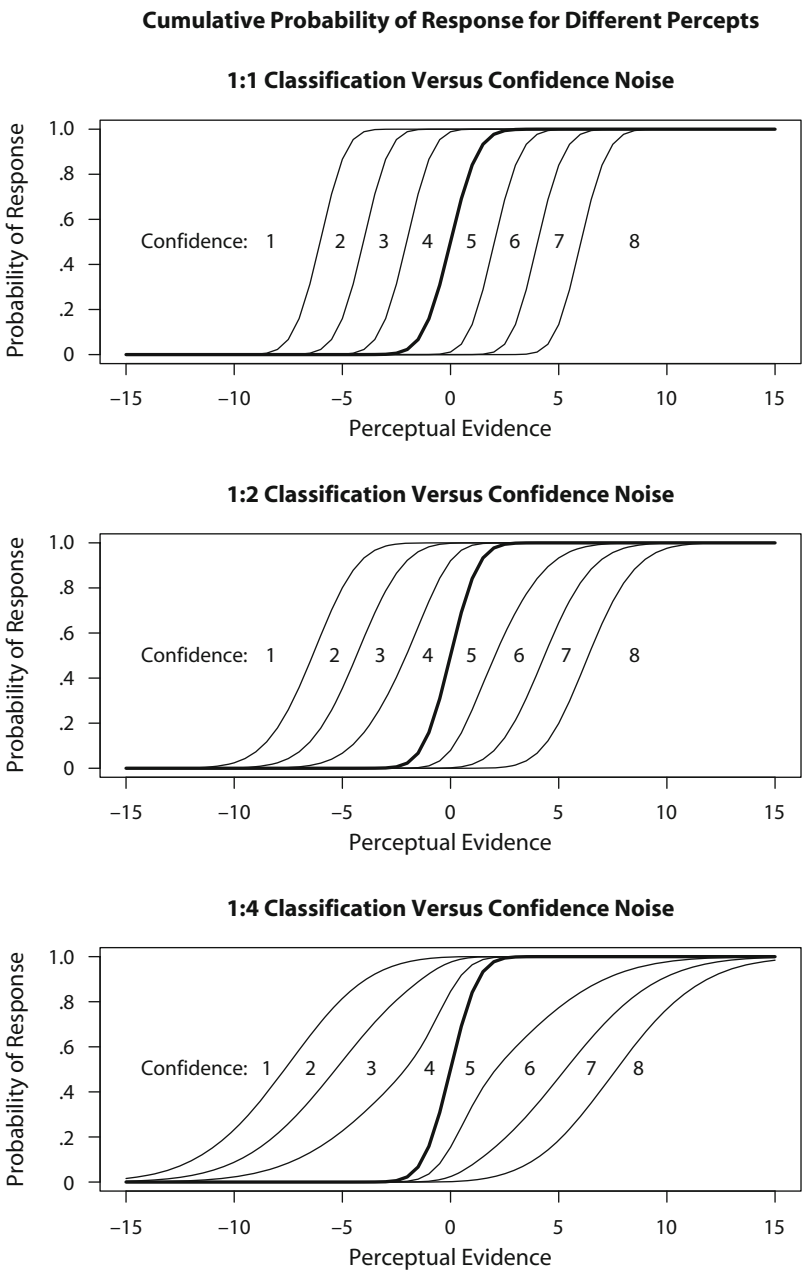

Figure 2. Depiction of the probabilistic mappings from percept onto confidence responses. For any specific level of perceptual evidence, the vertical distance between two lines represents the probability of producing each confidence rating. Classification noise is kept constant for the three panels (see the thick line), whereas confidence noise increases from the top to the bottom panel. The mean classification criterion was placed at 0, whereas means of the confidence criteria were ±2, ±4, and ±6.

sumed (for convenience) that on each trial, a classification criterion is sampled from a normal distribution and that a response class is determined on the basis of comparing the sampled percept to the sampled criterion. In order to produce a confidence rating, a similar process occurs within each response class. For an eight-level confidence scale in which four confidence classes occur for each response class, three criteria per response class are required. On each trial, positions of these confidence criteria are selected from normal distributions (by default with equal variance and means that are free parameters). In order to produce a confidence response, the model first examines the central classification criterion, and, depending on which side of the classification criterion the percept falls on, it samples the least confident confidence criterion in the proper direction. This conditional sampling continues until either a sampled criterion is found to be more extreme than the perceptual evidence, or no confidence regions remain. The

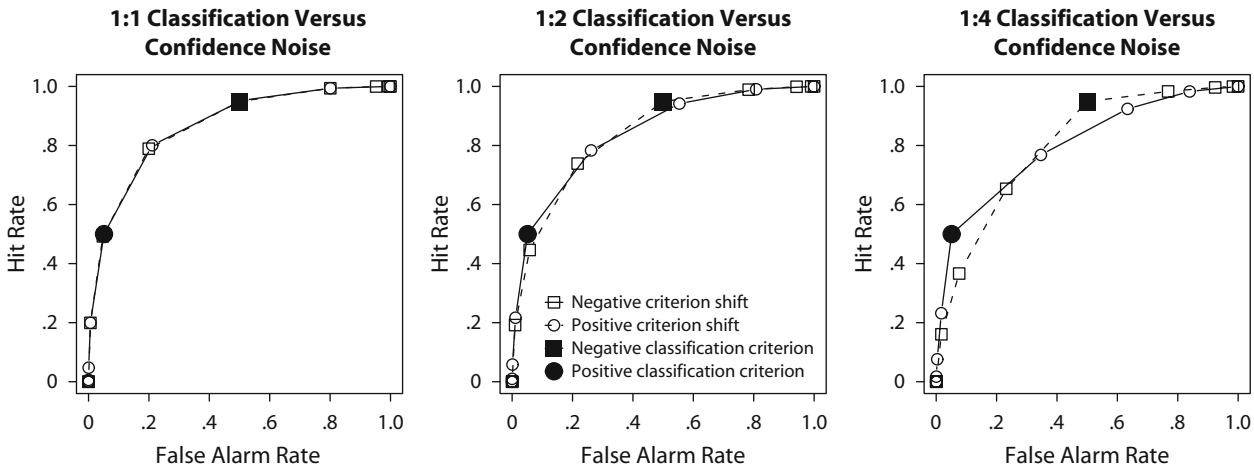

Figure 3. As confidence noise in the decision noise model increases relative to the constant classification noise, asymmetric C-ROC functions emerge. The left panel shows the equal noise condition, the center panel shows C-ROC functions with 1:2 ratios of criterion sampling standard deviation, and the rightmost panel shows C-ROC functions with 1:4 ratios.

confidence response is based on the position of the percept in relation to these sampled confidence criteria.[5]

Figure 2 shows three sets of response policies produced by the DNM that map percepts onto responses. Each panel shows a specific response policy determined by the mean and standard deviations of the decision criteria. For any specific level of perceptual evidence, the vertical distance between two adjacent functions indicates the probability of producing a specific confidence response. The top panel of Figure 2 shows a response policy in which the classification and confidence criteria have equal standard deviations (i.e., classification and confidence noise are equal); the middle panel shows a response policy in which the standard deviations of the confidence criteria are twice as large as that for the classification criterion, and the bottom panel shows a response policy in which the standard deviations for the confidence criteria are four times as large as that for the classification criterion. By comparing the three panels, one can see that confidence noise can be manipulated while maintaining the same level of classification noise (represented by the thick black line).

Using the DNM described so far, we can simulate data from signal detection tasks and examine the effects that unequal classification and confidence noise have on the resultant ROC and $U_R(k)$ functions. Doing so enables us to determine whether true criterion shifts could be detected in the presence of decision noise and the extent to which the proposed model can explain the observed crossover in the ROC functions (see Figure 1).

Predictions of the DNM

In order to show that decision noise can account for the crossover ROC functions, we performed a simulation with two normal stimulus distributions (A and B), with means of -2 and $+2$, and a standard deviation of 1 (simulating external noise). Furthermore, perception added normally distributed noise with a standard deviation of 2. Responses were formed by assuming that the classification criterion was drawn from a normal distribution with a mean of 0 and a standard deviation of 1. Confidence criteria had means of ± 2, ± 4, and ± 6, and standard deviations of either 1, 2, or 4 (depending on condition). These three conditions correspond to the three panels shown in Figure 2. For each condition, we examined both positive and negative criterion shifts by adding or subtracting 2 to the above criteria means.

Figure 3 shows how peaked and crossing C-ROC functions can be obtained if confidence noise is greater than classification noise. With equal levels of confidence and classification noise (left panel), two completely overlapping C-ROC functions are produced in response to criterion shifts. In this case, observed C-ROC functions lie along the same contour; thus, they cannot be used to discriminate between perceptual and decision noise. The middle and right panels show the obtained C-ROC functions as confidence noise increases with respect to the classification noise. As the ratio of classification noise to confidence noise changes from 1:1 to 1:4, peaked and crossing C-ROC functions emerge.

Although the distortions of the C-ROC function may be explained by decision noise, Balakrishnan's (1998a) measures of criterion shift may still be able to detect true shifts masked by decision noise. Consequently, we computed $U_R(k)$ functions on the simulated data (shown in Figure 4), examining three ratios of decision noise (one per panel) and considering three criterion shifts: no shift, a small shift (1 unit), and a large shift (2 units).

The $U_R(k)$ function can be used to detect criterion shifts by determining the criterion (k) at which the function peaks. Unshifted response policies should peak at $k = 4$ in our example. Results show that when decision noise is relatively small, true shifts in decision criterion can be detected using the $U_R(k)$ measure. However, with increases in confidence noise with respect to classification noise, these shifts become harder to detect. For the moderate confidence noise condition, the large shift can be detected, but the smaller shift produces estimates of $U_R(k)$ that are about equal for

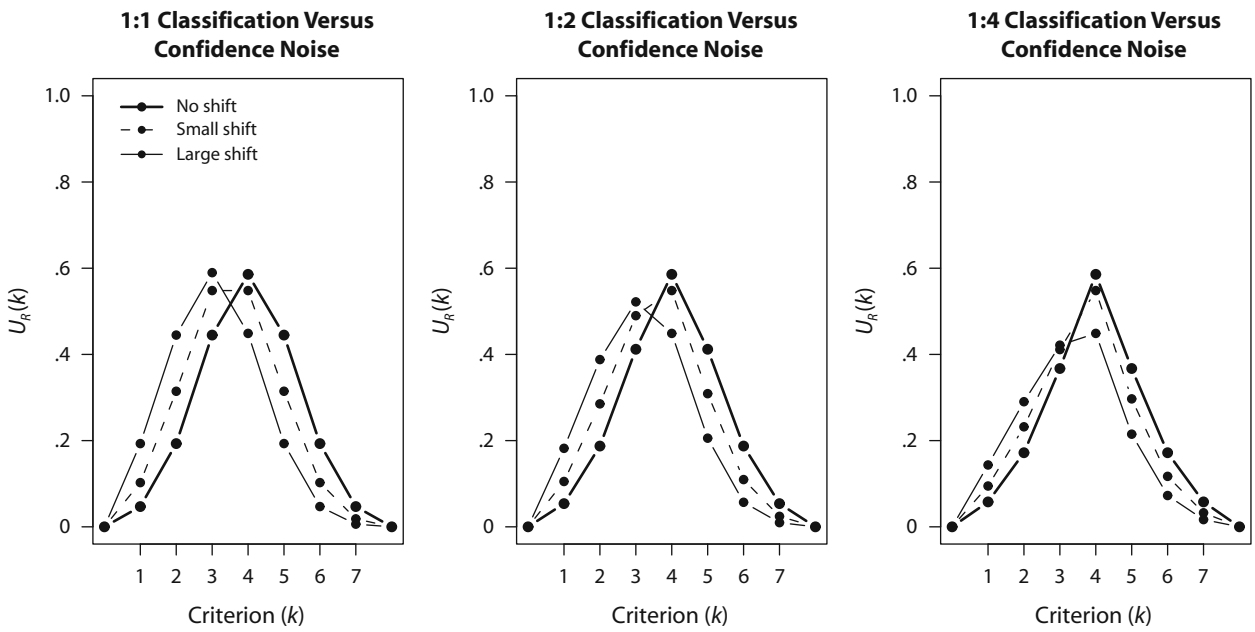

Figure 4. Predicted $U_R(k)$ functions produced by the decision noise model for different levels of noise and different criterion shifts (no shift; small shift, 1 unit; large shift, 2 units). As the confidence noise grows larger than the classification noise, the shift in the peak of the function $U_R(k)$ disappears, possibly explaining the fact that such shifts are rarely found in empirical data, despite changes in β.

the third and fourth criterion. As noise increases more, the small shift becomes undetectable, whereas the larger shift becomes ambiguous. This simulation demonstrates that if confidence noise is greater than classification noise, then a peak in the $U_R(k)$ function can appear at the medial confidence point, even if the decision criteria shift.

In addition to finding no measurable shift in the peaks of the $U_R(k)$ function, Balakrishnan (1998a, 1998b, 1999) observed suboptimalities in responding, which provide additional evidence against criterion shifts. These suboptimalities manifested themselves as low confidence A responses that were given more often in response to B stimuli than to A stimuli (or vice versa). This result indicates a suboptimal decision rule, because a movement of the decision criteria to optimal locations would eliminate such suboptimalities. As Treisman (2002) pointed out, the below-

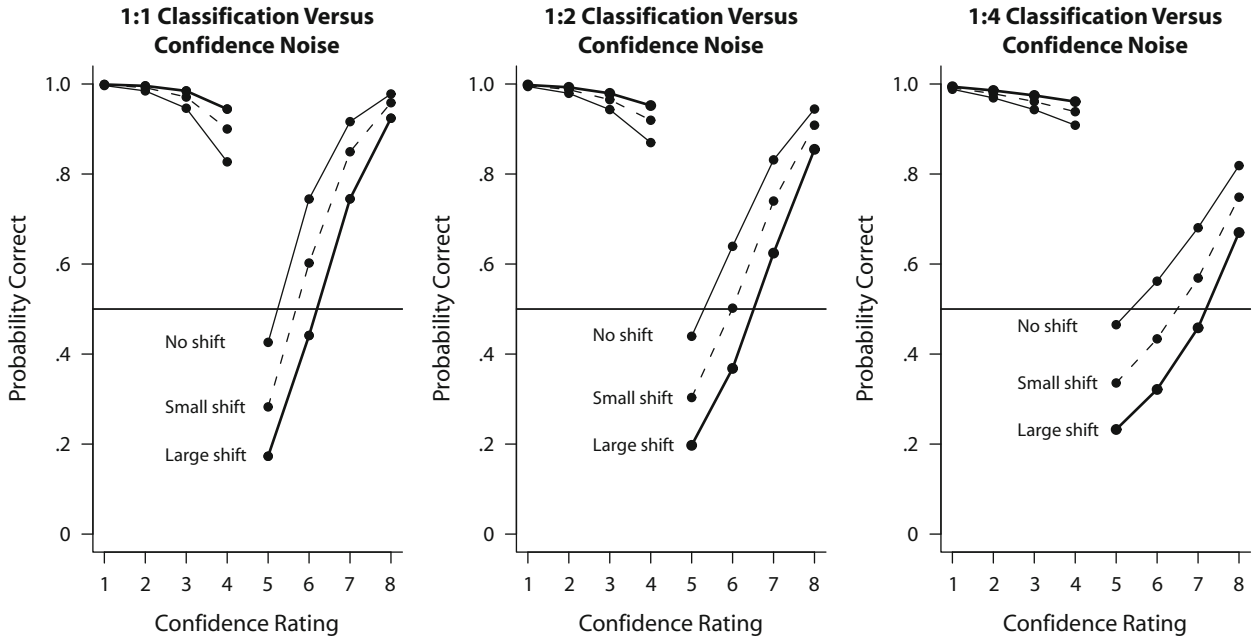

Figure 5. Predicted probability of correct responding for each confidence level produced by the decision noise model. Simulations for different levels of noise and criterion shifts (no shift; small shift, 1 unit; large shift, 2 units) are shown.

chance response accuracies for low confidence responses observed by Balakrishnan do not necessarily imply that the criterion has not changed at all; the shift may have just been smaller than optimal (see also Green & Swets, 1966, pp. 90–92; Healy & Kubovy, 1981). In our simulations of the DNM, we explored whether such a conservative shift can explain these suboptimalities even in the presence of decision noise. Just as is found in human data, the DNM should be able to produce results in which suboptimalities can be detected, but criterion shifts cannot.

Figure 5 shows simulated functions produced by the DNM for a base rate signal probability of .2. In each panel (representing increasing levels of confidence noise), overall accuracy for each confidence response is plotted. An optimal criterion shift would move all responses above .5; however, for the conservative criterion shifts simulated by the DNM, suboptimalities are detected in each of the decision noise conditions, despite the fact that the $U_R(k)$ functions in Figure 4 do not always reveal criterion shifts. This result demonstrates that it is, in principle, possible for decision noise to mask shifts in the criterion and produce suboptimal responding.

These simulations show how the findings of Balakrishnan (1998a, 1998b, 1999) can be explained in terms of decision noise. However, as was pointed out by Balakrishnan (2002), the peaked and shifted ROC functions and corresponding $U_R(k)$ functions could also have stemmed from changes in the perceptual distributions. In order to help distinguish between these two explanations, we will use a technique in which we add external noise to the stimuli and examine the relationship between distal stimulus and response category. In the next section, we provide the theoretical grounding for different types of ROC functions that help in making this distinction.

ROC Functions

In classic SDT, it is typically assumed that the perceptual distributions are normally distributed. The validity of this assumption for a specific data set is rarely tested, although some evidence for its validity can be obtained by examining an ROC function. Ideally, an ROC function is formed by manipulating the observer's classification criterion across many different conditions. However, this is data intensive and time consuming, because it requires testing under numerous base rate or payoff conditions. Consequently, researchers often use a more efficient method based on confidence ratings. Doing so allows an ROC function to be computed for a single experimental condition.[6]

Following Zhang, Riehle, and Requin (1997), in this section we will review different types of these single-condition ROC functions that can be used to make inferences about the underlying processes involved in a signal detection task. In particular, we will distinguish between stimulus-related and response-related ROC functions. Figure 6 schematically depicts the different sources of data that may be available during a signal detection task. In the diagram, binary categorical variables are represented by rectangles, whereas multilevel ordinal-scale variables are represented by ovals. The central flowchart shows how a nominal stimulus is selected and a distal stimulus is produced, which is then perceived by the observer, who produces a response. Along with a classification response of the observer, recordings can be made of the distal (presented) stimulus, activation levels or firing rates in the neural tissue of the observer, or response-related variables, such as response time (RT) or subjective confidence level. ROC functions can be computed by pairing a binary classification variable (e.g., stimulus or response category) with a multilevel ordinal-scale variable (e.g., firing rate or confidence).

Two classes of ROC functions computable from neural recording data were defined by Zhang et al. (1997): the stimulus-related and response-related ROC functions. We will first discuss stimulus-related ROC functions, which are commonly used to make inferences about the shapes and the variances of the perceptual distributions. Then, we will show how response-related ROC functions can be constructed to make inferences about the mapping from perception to response.

Stimulus-related ROC functions. A commonly used stimulus-related ROC function is the confidence-based ROC (C-ROC) function, which is computed by calculating hit rate and false alarm rate for each transition between confidence states. The steps involved in constructing a stimulus-related C-ROC function are shown in the top row of Figure 7, and a detailed example is given in Appendix B. First, one administers a signal detection task experiment, collecting confidence ratings. The data are divided into two classes: signal trials and noise trials, and confidence rating distributions are formed for each class (leftmost panel). These empirical distributions are typically taken as estimates of the underlying signal and noise distributions, although decision noise can distort this relationship. Next, one computes the cumulative density function of the signal and noise distributions (center panel). The function $U_R(k)$ is computed by taking the difference between these two functions, and the ROC function is computed by plotting the cumulative distribution function (subtracted from 1.0) of the signal and noise distributions against one another, for all transitions between confidence responses (rightmost panel). In this figure, we plot these functions as smooth lines, even though they are empirically derived from dozens of discrete points along the cumulative density functions (we will do the same in later figures whenever more than eight points contribute to a function). The shape of the resulting C-ROC function is determined by the shape of the perceptual distributions and response policies. For chance performance, the C-ROC function approximates the main diagonal. To the extent that the observer is able to discriminate between signal and noise trials, the C-ROC function becomes convex. Therefore, the area under the C-ROC function can be used as a measure of the relationship between stimulus class and confidence: The upper bound on the area under the ROC function is 1.0, and chance performance corresponds to an area of .5. In the example, the signal distribution has less variability than the noise distribution; thus, the obtained ROC function is not symmetric around the negative diagonal.

Other measures besides confidence have been used to form stimulus-related ROC functions. These include RTs

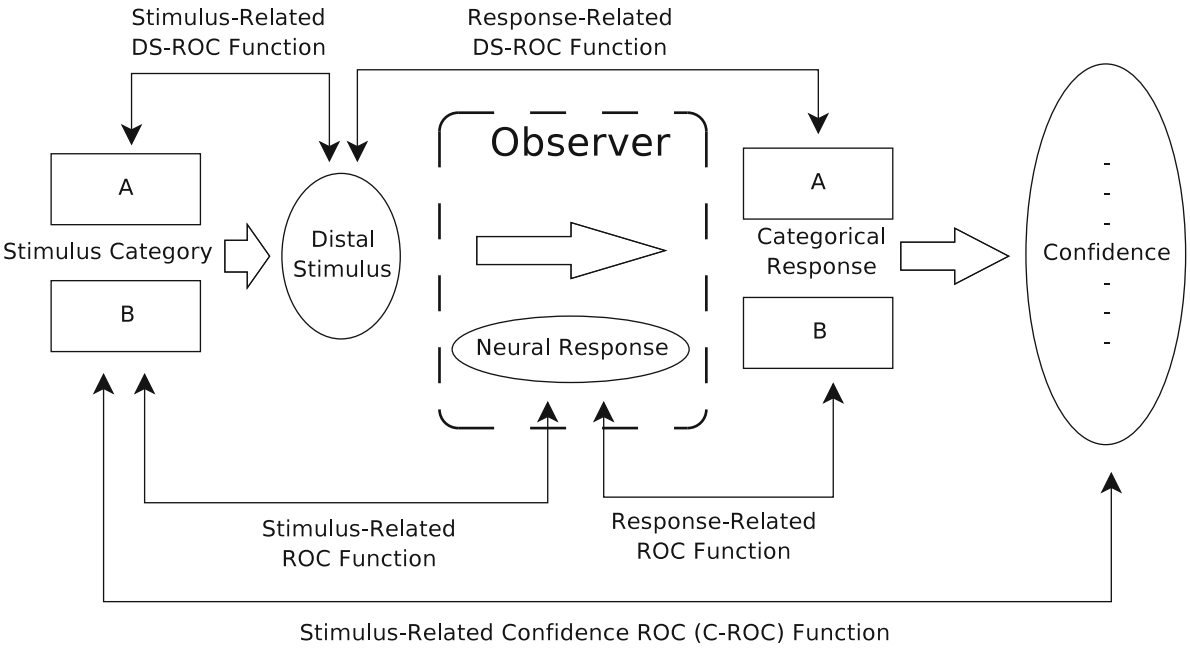

Figure 6. Schematic diagram of data sources available for measurement during the signal detection task. Binary categorical variables are shown as rectangles; multilevel ordinal-scale variables are shown as ovals. ROC functions can be formed by examining an ordinal-scale variable conditioned on a categorical variable.

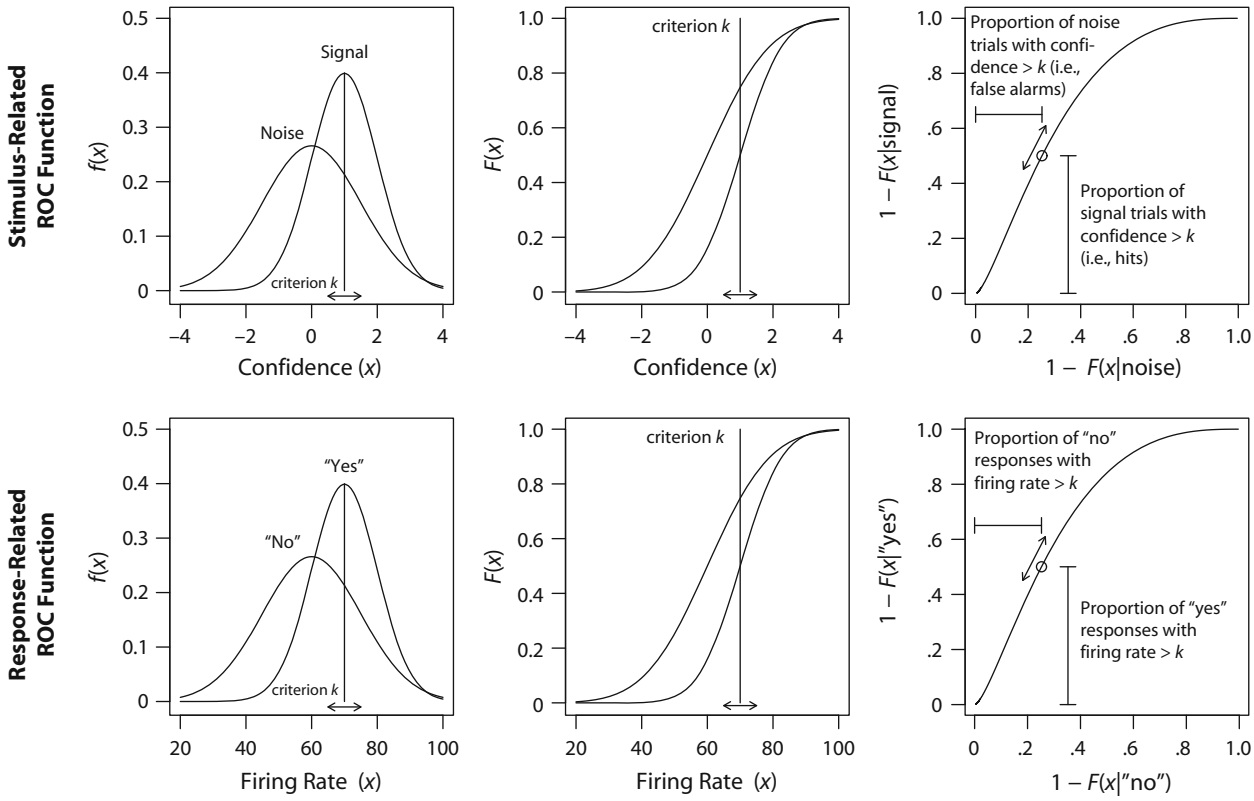

Figure 7. Steps in the construction of stimulus-related and response-related ROC functions. First (left panels), trials are sorted into two categories (either by stimulus or response class), and relative frequency histograms (*f* [*x*]) are formed representing the distributions over a secondary variable *x* (such as the firing rate of neurons or confidence estimates). These histograms are converted to cumulative frequency distributions (*F*[*x*], middle panels), and the corresponding cumulative frequencies are plotted against one another for every possible criterion *k*, along the secondary variable *x*, forming the corresponding ROC functions (right panels). A worked-out example of this process is given in Appendix B.

and neural recordings (see, e.g., Newsome, Britten, & Movshon, 1989; Zhang et al., 1997), which have been used to examine the discriminability of individual cells in a monkey cortex during perceptual-motor tasks. Other factors can be used as well, including neural activity in a brain imaging context, skin conductance, pupil dilation, and even data from the environment (such as properties of the stimulus or ambient noise). These ROC functions will only reveal aspects of an observer's sensitivity if the dependent measure is related to the observer's ability to discriminate between the stimulus classes. If there is no relationship, then the resulting ROC function will approximate the line $y = x$.

Even though stimulus-related ROC functions based on properties of the environment provide no information about the observer, some can still be useful. For example, suppose that a stimulus is presented in noise so that it is impossible to perfectly discriminate signal trials from noise trials. If one uses an independent measure of the intensity of the distal stimulus to form a stimulus-related ROC function (we call this a *stimulus-related distal stimulus ROC function*, or *stimulus-related DS-ROC* for short), then the function traces out an upper bound on the accuracy that could be attained in that experiment. Similar procedures that introduce external noise have a long tradition in the investigation of decision rules in SDT (see, e.g., Kubovy & Healy, 1977; Lee, 1963; Lee & Janke, 1964; Lee & Zentall, 1966), and such techniques have been adopted more recently to investigate attentional mechanisms (Lu & Dosher, 1998) and to identify features that humans are using in visual perception tasks (Gold, Murray, Bennett, & Sekuler, 2000). Despite the limited insights that a stimulus-related DS-ROC function can provide, a similar function computed on the response classes can be useful and reveal the relationship between the stimulus and response category. We will discuss such response-related ROC functions next.

Response-related ROC functions. Zhang et al. (1997) demonstrated how the functional locus of a cell along the perception–decision–response continuum can be isolated by forming response-related ROC functions in addition to the stimulus-related ROC functions discussed previously. The lower panels of Figure 7 show how a response-related ROC function can be computed by examining the response class in comparison with a secondary variable. As an illustrative example, consider a yes–no categorical decision in which the neural firing rate of some area of cortex is recorded on each trial. First, the data are sorted into two classes corresponding to "yes" and "no" responses, and distributions of neural firing rate are computed for each response class (left panel). Next, the cumulative density functions for these two distributions are formed (center panel). Finally, for each firing rate, the cumulative densities for "yes" and "no" responses are subtracted from 1.0 and plotted against one another to form the response-related ROC function (rightmost panel). This function shows the relationship between the measured variable and the categorical response. As in the stimulus-related ROC function, if the measured variable is unrelated to the measured response, then the function will approximate the line $y = x$.

This function does not map out accuracy directly, because no information about the stimulus category is used. Instead, it maps out the relationship between the categorical "yes" or "no" responses and the underlying firing distributions of a neuron. Therefore, just as traditional stimulus-related ROC functions effectively answer the question, "If confidence level k had been used as the classification criterion, what would accuracy have been for each *stimulus* class?" response-related ROC functions effectively answer the question "If firing rate k were the objective division between the two stimulus classes, what would accuracy have been for each *response* class?" Thus, in contrast to the typical stimulus-related ROC functions, which plot the proportion of hypothetical "yes" responses for the two stimulus classes, response-related ROC functions plot the proportion of hypothetical signal trials for "yes" responses against those for "no" responses.

Although originally defined to examine the relationship between a firing neuron and an overt response, the same computation can be performed on any multivalue ordinal-scale variable to determine the relationship between that variable and the categorical response. Philiastides and Sajda (2006), for example, related neural components obtained from scalp EEG recordings to responses. Another example is the response-related ROC function based on the (noisy) distal stimulus. This function measures the relationship between the presented stimulus intensity and the response by evaluating the cumulative predictive sensitivity along all the presented intensities. We refer to this as the *distal stimulus ROC* (DS-ROC) function, or the *response-related DS-ROC* when the meaning is ambiguous. We present a sample calculation of C-ROC and DS-ROC functions in Appendix B.

The DS-ROC can be useful, because it measures the relationship between the stimulus and the classification response across a range of stimulus values without making use of confidence ratings. Thus, if the DS-ROC shows the asymmetries observed in the C-ROC functions, then we can infer that at least part of the asymmetry stems from processes prior to confidence assessment and in the mapping from distal stimulus to percept. However, if only the C-ROC functions show the asymmetry, then we can isolate the asymmetry to processes that occur in the mapping from percept to confidence response. Thus, we may be able to use DS-ROC functions to distinguish between two alternative accounts of the crossover in the C-ROC functions.

The previous simulations showed how—when decision noise was large—the DNM produced data in which C-ROC functions crossed, and for which the $U_R(k)$ and suboptimality functions could not detect true shifts in a decision criterion. We will next present simulations that show how C-ROC and DS-ROC functions respond to changes in perceptual distributions and decision policies. We simulated C-ROC and DS-ROC functions for four variations of the DNM, factorially manipulating base-rate dependent perceptual noise and the ratio of the noise associated with the classification and confidence criteria (shown in Figure 8). These simulations were identical to the previous ones, except that we only considered two ra-

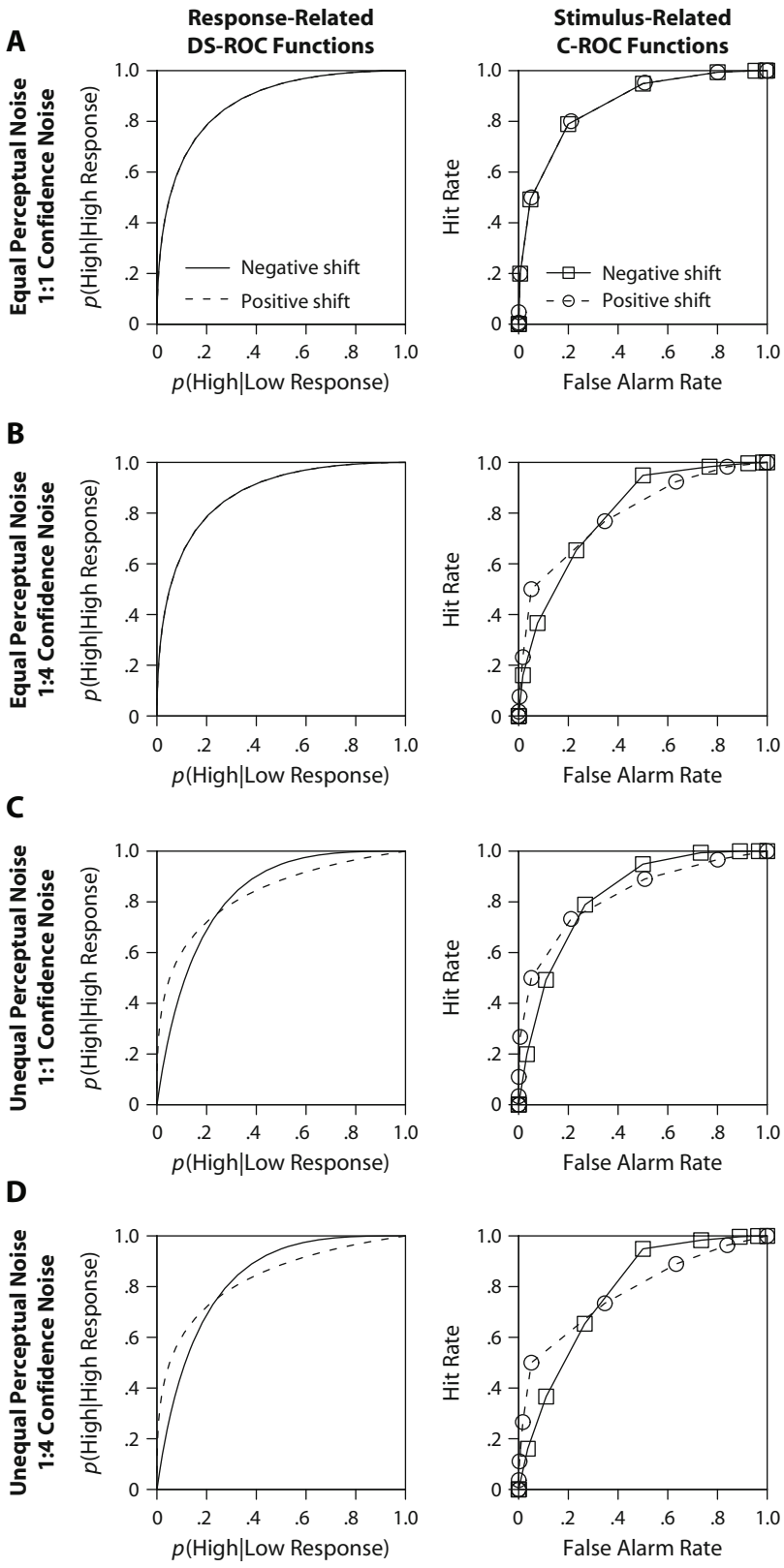

Figure 8. Predictions of the decision noise model for different sources of noise. Differences in the perceptual distribution variance that depend on base rate/payoff produce crossover DS-ROC and C-ROC functions. If true criterion shifts exist, then crossover C-ROC functions will be obtained if the variances of the confidence criteria are larger than the variance of the medial classification criterion.

tios of classification and confidence noise: 1:1 and 1:4. Furthermore, we considered two perceptual noise regimes: one in which signal and noise distributions had equal variance (i.e., equal perceptual noise), and one in which the ratio between noise and signal standard deviations was 2:3 for positive criterion shifts and 3:2 for negative criterion shifts (i.e., unequal perceptual noise). The DS-ROC was formed using the simulated stimulus intensity. Each panel shows two ROC functions corresponding to positive and negative shifts in the response policy (as is often assumed to occur with changes in base rate or payoff). In all cases, the SDT statistic β was sensitive to these shifts.

Panel A of Figure 8 shows simulated DS-ROC and C-ROC functions with equal perceptual noise (i.e., the variance of the signal and noise distributions were the same and did not depend on condition) and equal decision noise (i.e., the variance of the classification criterion was the same as the variance of the confidence criteria). ROC functions for the two criterion shift conditions lie on top of one another, producing results indistinguishable from those expected by classic SDT. This occurs even though decision noise is present, and, in a typical analysis, this noise would have been incorporated into the overall estimate of sensitivity. Panel B shows simulated ROC functions with equal variance for signal and noise distributions, but confidence noise that was greater than classification noise. Here, asymmetries are observed for only the C-ROC function, because the source of the asymmetry is in the decision process. Data showing this pattern are especially diagnostic, because they imply that the asymmetries arise during decision rather than during perceptual processes. Panel C shows the effect when only the variance of the perceptual distributions differ (i.e., the variance of signal and noise distributions depend on base rate): It produces the crossover in the DS-ROC functions, and this crossover carries over to the C-ROC functions. Finally, combining the two sources (panel D) shows asymmetries in both ROC functions as well. Results such as those in panels C and D are not easily distinguishable, and if empirical data show crossovers for both DS-ROC and C-ROC functions, then it would be difficult to determine the contribution of decision noise to the effect.

The demonstration in Figure 8 shows that if the perceptual distributions change shape in response to base rate manipulations, then both the DS-ROC and the C-ROC functions change shape. In contrast, if the mapping between percept and response becomes uncertain because of decision noise, then the C-ROC function changes shape, but the DS-ROC does not. In the next section, we will report an experiment that we conducted to demonstrate how this type of analysis can be used to determine the source of the crossover C-ROC functions discussed earlier.

EXPERIMENT
The Effects of Perceptual and Decision Noise

In order to demonstrate how the locus of asymmetries in an experiment can be identified, we conducted a stimulus classification experiment involving external noise. Doing so enabled us to form DS-ROC functions that were based on actual stimulus intensities and to demonstrate the ability of the DNM to fit these data.

Method

Participants. Participants were 49 undergraduate students from Indiana University who received partial completion of course credit in exchange for their involvement in the experiment.

Procedure. Each participant took part in a single experimental session that lasted less than an hour. Participants sat in a private dimly lit booth approximately 30 cm from a 17-in. CRT display screen, which was attached to a desktop PC and set to a resolution of 800 × 600. Stimulus presentation and response collection was performed using a custom-written experiment script managed by the PEBL Version 0.05 experiment language interpreter (Mueller, 2005). Stimuli were presented in 12-point Bitstream Vera Mono typeface.

On each trial, participants were shown a 10 × 10 character grid that contained randomly distributed "∗" symbols. Their task was to determine which category the stimulus came from: "A" (which contained, on average, 46 symbols) or "B" (which contained, on average, 54 symbols). The actual number of symbols on each trial was determined by sampling from a normal distribution whose mean depended on the stimulus class and whose standard deviation was 5. Participants performed 24 blocks of this task under different base rate and response conditions. For half of the blocks, participants responded either "A" or "B" by pressing one of two keys on the keyboard (the "Z" and "/" keys); the other blocks involved a confidence rating task in which participants were instructed to give a rating of 1, 2, 3, or 4 to indicate A stimuli and 7, 8, 9, or 0 to indicate B stimuli (by pressing the corresponding number keys along the top of the keyboard). In the confidence task, ratings given with the index fingers (4 and 7) indicated low-confidence responses, and responses given with the little fingers (1 and 0) indicated high-confidence responses. Blocks of stimuli were presented under three different base rate ratios: 2:1, 1:1, and 1:2. Payoff regimens were designed to encourage the use of high-confidence responses only when participants were very certain of their classifications: In order from *least confident* to *most confident*, rewards of 1, 2, 3, or 4 points and losses of 1, 3, 5, and 7 points were given. All blocks from each base rate were contiguous, and, within each base rate condition, the confidence and forced-choice task alternated. In contrast with the experiments reported by Balakrishnan (1998a) in which targets appeared on approximately 10% of trials, our base rate manipulations were fairly small (i.e., 2:1 and 1:2). These were chosen to allow for a substantial number of trials in the low base rate condition while keeping the duration of the experiment under 1 h, because experiments with extremely rare stimuli can be contaminated by effects stemming from attentional lapses and the failure to maintain vigilance (see, e.g., Loh, Lamond, Dorian, Roach, & Dawson, 2004). In total, four 60-trial blocks were presented for each task and each base rate condition, for a total of 1,440 trials per observer.

Initial data processing. Although a point-based payoff was used in the task, no monetary reward was given, so we anticipated that some participants would not be engaged in the task or would otherwise fail to perform according to instructions. An initial examination of individual participants' data showed that 6 participants scored substantially fewer points than others or used primarily two confidence responses. As a result, only data from the 43 remaining participants were analyzed further. Our use of external noise creates ambiguity in the stimulus class designation, because the nominal stimulus class (by which base rate varied and upon which feedback was given) was impossible to discriminate perfectly. However, even when a stimulus nominally arose from one distribution, it may have been more likely to have come from the other distribution. Thus, we computed all statistics that required specifying stimulus class (e.g., hit rate, false alarm rate, etc.) on the basis of the ideal stimulus category for each distal stimulus: Stimuli with fewer than 50 symbols were designated

as A stimuli, and those with 50 or more symbols were designated as B stimuli. Finally, the first block of trials in each task and base rate condition (a total of six blocks) was considered practice and was not included in the analysis. In order to simplify the presentation of results, we will refer to the confidence responses made by pressing the 7, 8, 9, and 0 keys as 5, 6, 7, and 8, respectively, so that the responses range from 1 to 8 with no gaps.

Results

SDT statistics. We began by computing the traditional SDT measure of d' and β for each participant. For the forced choice task, mean d' values were 0.85, 0.85, and 0.92; and mean β values were 0.956, 1.065, and 1.19 for the 2:1, 1:1, and 1:2 base rate conditions, respectively. For the confidence ratings task, the corresponding mean d' values were 0.91, 0.88, and 0.97; the mean β values were 0.92, 1.02, and 1.12 for the 2:1, 1:1, and 1:2 base rate conditions, respectively. Individual estimates of β and d' were submitted to separate ANOVA procedures to determine the effects of base rate condition and test type, treating participant as a randomized factor. Results showed that neither base rate [$F(2,84) = 2.2$, $p > .1$] nor test type [$F(1,42) = 1.8$, $p > .1$] reliably affected d', although both base rate [$F(2,84) = 35.8$, $p < .001$] and test type [$F(1,42) = 71.5$, $p < .001$] reliably affected β. The reliable shifts in β were in the directions expected by SDT in response to base rate manipulations; the reliable effect of test occurred because β values were slightly higher for all base rate conditions in the forced choice task than in the confidence ratings task.

ROC functions. Next, we calculated C-ROC and DS-ROC functions from data pooled across participants. The C-ROC functions were formed for different base rate conditions of the confidence rating procedure. DS-ROC functions were formed from each base rate of the forced choice data as well as from each condition of the confidence rating procedure, mapping confidence responses into two response categories: "A" and "B." The DS-ROC was formed by computing the cumulative distributions across the number of presented stars in the display on each trial. The results are shown in Figure 9. The DS-ROC functions were nearly identical for response conditions, indicating that the confidence rating procedure did not affect the overall response category substantially. Furthermore, the DS-ROC functions were nearly identical for all base rate conditions, indicating that similar perceptual distributions occurred for each condition. However, we observed C-ROC functions that crossed, just as had been found previously by Balakrishnan (1998a, 1998b, 1999; see also our Figure 1). The changes in C-ROC functions that we observed are smaller than those shown in Figure 1, but it is important for one to keep in mind that our base rate manipulation (which was 1:2) was considerably weaker than that used by Balakrishnan (1998a; which was 1:9).

To test statistically whether two ROC functions differ is a challenge, because we are interested in whether the overall shape changes, independent of the positions of the individual confidence points. One way to do this is to transform the data into a space in which the functions are linear and to use standard statistical techniques to determine whether the slopes of the functions depend on base rate condition. The pooled ROC functions from our data set were essentially linear when transformed into z coordinates (i.e., the corresponding values of the standard normal distribution), although the C-ROC functions tended to have a small discontinuity at the transition between "A" and "B" responses. Consequently, in order to determine whether base rate condition had an effect on the shape of the ROC functions, we computed the slope of the z-transformed ROC functions for each participant. We then examined the mean z-transformed C-ROC and DS-ROC function slopes for the forced choice and confidence rating conditions of the experiment. Mean slopes for the C-ROC functions were 1.029, 0.934, and 0.906 for the 2:1, 1:1, and 1:2 base rate conditions, respectively. An ANOVA treating participant as a randomized factor confirmed that these differences were reliable [$F(2,82) = 5.57$, $MS_e = 0.031$, $p = .005$],[7] indicating that the C-ROC functions did indeed differ. In contrast, the slopes of the DS-ROC functions were not reliably affected by the base rate condition. For the confidence-rating procedure, mean z-transformed DS-ROC slopes were 1.01, 1.00, and 1.00 for the 2:1, 1:1, and 1:2 base rate conditions, respectively, which were not reliably different [$F(2,84) = 0.08$, $MS_e = 0.020$, $p = .92$]. For the forced choice procedure, mean z-transformed DS-ROC slopes were 1.01, 1.02, and 1.02 for the 2:1, 1:1, and 1:2 base rate conditions, respectively, which were also not reliably different [$F(2,84) = 0.019$, $MS_e = 0.02$, $p = .98$]. These results support our hypothesis that the violations of SDT that we observed stemmed from decision-related processes.

These results also suggest that confidence judgments are subject to more noise than are classification responses, and this decision-related noise may account for many empirical violations of SDT. Such noise could stem from a number of sources, such as learning, forgetting, maintenance of consistency, or response perseveration. Many of these accounts predict trial-by-trial dependencies, which we will examine next.

Response dependencies. Sequential dependencies between consecutive stimuli or responses have frequently been found in perception and detection tasks (Jones, Love, & Maddox, 2006; Parducci & Sandusky, 1965; Sandusky, 1971; Treisman & Williams, 1984; Ward, 1973). Trial-by-trial conditional effects could occur for a number of reasons, and they may be one important source of the decision noise found in our task. The conditional dependencies of different responses are shown in Figure 10. In this figure, each panel represents a different base rate condition from our experiment. The size of the circle at each point in the grid is monotonically related to the number of total confidence ratings in that condition across the entire experiment.

Separate χ^2 tests computed for each base rate condition showed that the dependency matrix deviated reliably from the null model (i.e., the hypothesis that the joint distribution of cell counts is the product of the row and column marginal distributions). Consequently, results showed reliable and strong trial-by-trial contingencies [$\chi^2(49) = 2{,}559, 2{,}903, 2{,}853$ for the conditions with more A stimuli,

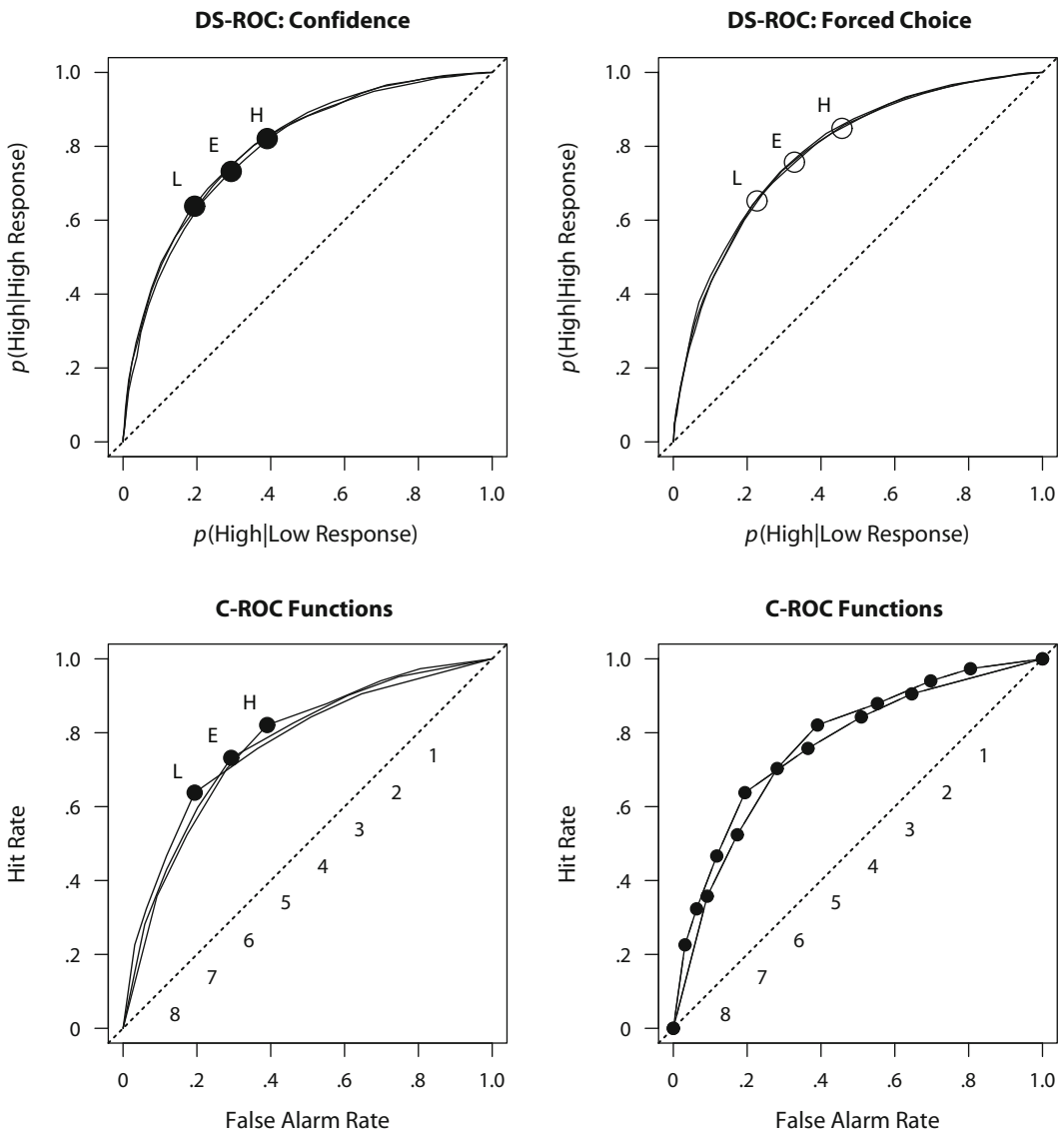

Figure 9. DS-ROC and C-ROC functions computed by pooling data from all participants. The bottom left panel shows all three base rate conditions (filled circles show classification criteria), and the bottom right panel shows only the 2:1 and 1:2 base rate conditions with all confidence criteria. C-ROC functions were formed by considering ideal rather than nominal stimulus categories. Numbers along the diagonal represent the approximate regions of the different confidence responses. L, low stimuli more frequent; E, high and low stimuli equally frequent; H, high stimuli more frequent.

equal A and B stimuli, and more B stimuli, respectively; each $p < .0001$]. There is a strong tendency to repeat responses (the positive diagonal), with a lesser tendency to reflect to the corresponding confidence for the opposite response category (the negative diagonal). If no conditional responses were occurring, then each column should have shown the same relative pattern, which clearly did not happen. Some of the repetition and reflection may stem from individual differences in response strategies, with some participants focusing, for example, on the two highest confidence ratings, and others focusing on the two lowest confidence ratings. However, similar patterns emerge when we restrict the analysis to just the participants who distributed their confidence responses over all confidence levels. Nonetheless, this provides a second important demonstration that suboptimalities in response processes are important contributors to performance in our task, and presumably in other tasks as well.

Criterion shifts. We also assessed our data using several measures introduced by Balakrishnan (1998b). These results are shown in the top row of Figure 11, with error bars indicating ± 1 standard error. As we reported earlier, the traditional SDT measure β differed reliably across base rate conditions. Yet the peak of the $U_R(k)$ functions (left panel, top row) did not change in response to base rate, indicating no criterion shift. Next, we examined the probability of correct response for each confidence level (second column, top row). Here, our results showed that par-

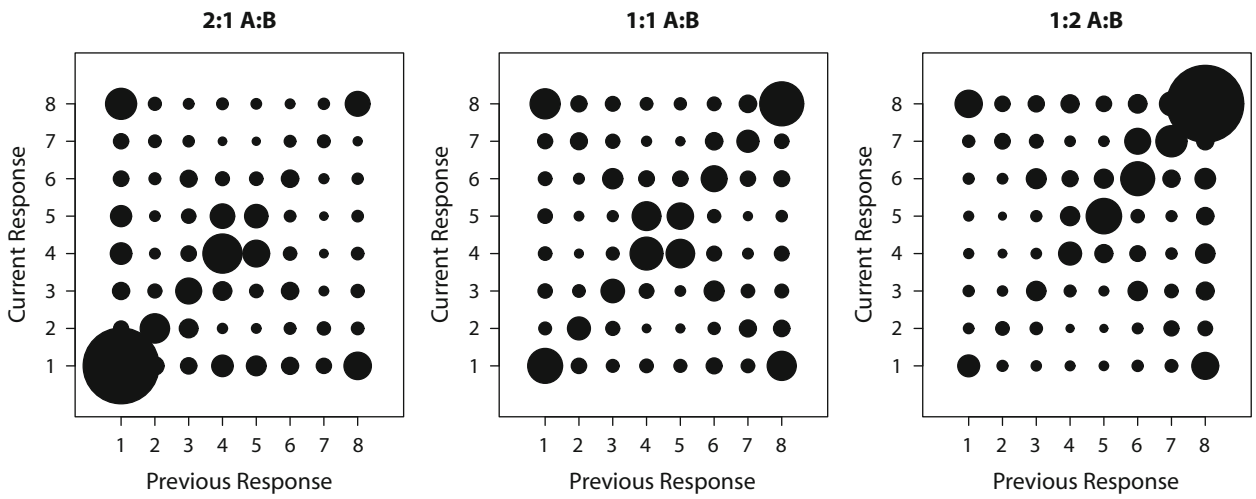

Figure 10. Trial-by-trial dependencies from the confidence-rating task. Each column represents the previous response, and each row represents the current response. The size of filled circles is monotonically related to the number of trials in each conditional category.

ticipants were, on average, above chance for all confidence responses, a result that is consistent with a criterion shift.

Balakrishnan (1998b) also introduced a distribution-free measure of bias, Ω, which indexes the proportion of biased responses. As he pointed out, an upper bound on this measure can be computed on the basis of the response proportions for different confidence levels. In particular, low proportions of responses for the confidence levels flanking the peak of the $U_R(k)$ function correspond to a low upper bound of this bias index. In our experiment, a substantial proportion of responses fell into the low confidence regions; therefore, we are unable to exclude the possibility that a sizable proportion of responses were biased, even though the peak of the $U_R(k)$ function of our data remained in the central position for all base rate conditions. As we will show below, however, our model can account for the central peak in the $U_R(k)$ function, even when the upper bound of biased responses is low.

So far, our results have supported the decision noise hypothesis, demonstrating changes in β, C-ROC functions that cross without any effect on the DS-ROC functions, and substantial trial-by-trial response dependencies. If the model does provide a reasonable account of the processes involved in the signal detection task, then we should be able to provide precise quantitative fits to data as well. We will demonstrate such fits in the next section.

DNM FITS TO DATA

The DNM is designed as a simple extension of SDT to capture decision noise. It does not directly incorporate detailed accounts of decision uncertainty like the trial-by-trial dependencies we found in our present experiment (although more complex models do; e.g., Treisman & Williams, 1984). Nevertheless, we attempted to determine whether the model could produce results like those seen in our present experiment and account for similar empirical phenomena that arise across a diverse set of experimental paradigms (visual detection, visual classification, and recognition memory).

In order to simulate data, we assumed that the means of confidence criteria were symmetric and fixed relative to the classification criterion, even when that criterion was shifted. We assumed that performance was affected by perceptual noise, classification noise (i.e., the variance of the classification criterion), and confidence noise (i.e., the variance of the other criteria). In practice, the obtained fits tended to minimize perceptual noise, so we eliminated perceptual noise for these simulations. This result probably does not indicate that there was no noise in stimulus perception, but it is likely to be a consequence of perceptual and decision noise trading off and therefore being difficult to separate.[8]

This model has a total of $(n - 1) + 5$ free parameters—where n indicates the number of confidence states allotted to each response category (in the present experiment, four)—and, because of the trade-off between variance sources, we use only $(n - 1) + 4$ of these, bringing the total number of free parameters in our experiment to seven. The fitted values and descriptions of these parameters are found in Table 1. Even the simplest extension of SDT to confidence responses would likely require $(n - 1) + 3$ free parameters (i.e., $n - 1$ confidence criteria, a central criterion, a shift in response to base rate, and a noise parameter); thus, our model adds either one or two parameters that control decision noise.

In order to simulate data, we estimated the perceptual distributions and decision policies numerically rather than via a Monte Carlo simulation of the entire perceptual process. A detailed description of this approach is provided in Appendix A.

Application of the DNM to the Present Experiment

By adjusting the parameters of the DNM, we attempted to fit the following measures on data: (1) $U_R(k)$ functions that do not detect shifts in criterion, (2) probability

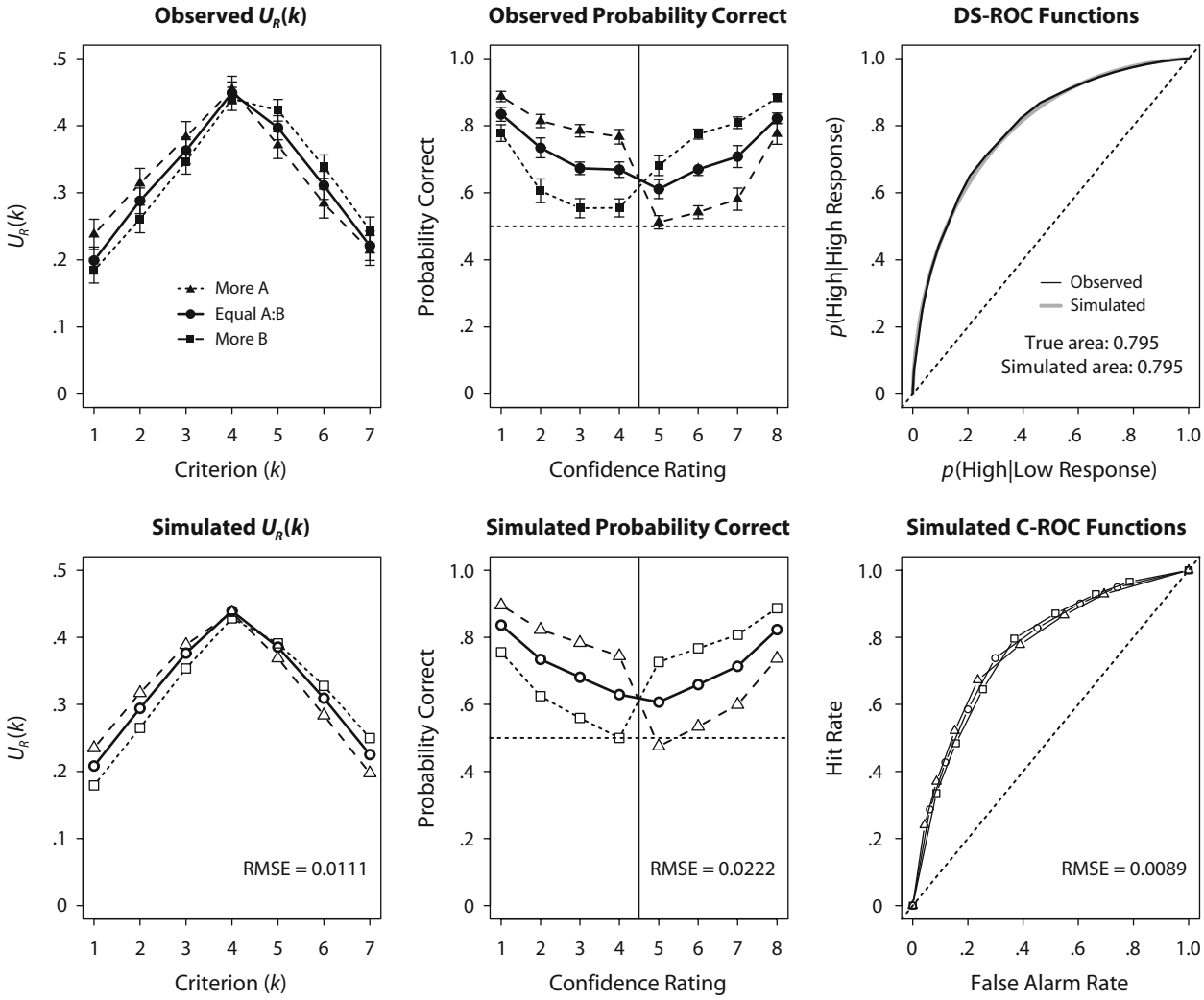

Figure 11. Statistics indicating criterion shifts for data and model, using ideal stimulus categories. The first column shows observed and predicted $U_R(k)$ function for experiment and simulated decision noise model. The second column shows probability correct. The third column shows observed and simulated DS-ROC functions (top panel) and simulated C-ROC functions (bottom panel), which can be compared with the corresponding observed functions in Figure 9. Observed data are shown with filled symbols; simulated data are shown with open symbols. Error bars show ±1 standard error.

correct functions that produce close-to-chance accuracy for low-probability/low-confidence responses, (3) identical DS-ROC functions for different base rates, and (4) the crossover C-ROC functions for three base rates, whose midpoints are the overall hit and false alarm rate for each condition. We estimated the best-fitting parameters by minimizing the sum of the root mean squared error (RMSE) badness-of-fit statistic for $U_R(k)$, probability correct, and C-ROC functions, and the absolute difference in the observed and simulated areas under the DS-ROC functions. Estimates were obtained using a Newton-type gradient descent method as implemented in the `nlm` function of R (R Development Core Team, 2006), starting from numerous initial positions in order to minimize the chance of inappropriate local minima. The estimated functions resulting from these simulations are shown in Figure 11, and the obtained parameter values are in Table 1. Parameters were all scaled with respect to a stimulus intensity unit corresponding to a single "∗" symbol.

Our results show that we were able to reproduce the major results of our experiment with a relatively simple model. Our model had a total of seven free parameters, which we used to fit the C-ROC functions, the overall shape and area of the DS-ROC function, and the overall level of performance in the probability correct functions. $U_R(k)$ is determined by the C-ROC function and does not represent any additional degrees of freedom. Because probability correct is a function of the same response distributions but incorporates the base rate, it represents only three additional degrees of freedom in the data. Notably, the model behaves like the human observers: It does not show changes in the peak of $U_R(k)$, and it produces shifts in β and the crossover C-ROC functions while also generating identical DS-ROC functions. However, there

Table 1
Parameter Estimates Obtained by Fitting the Decision Noise Model to the Data in the Present Experiment and to Those of Balakrishnan (1998a)

		Balakrishnan (1998a)	
Parameter	Present Experiment	Experiment 1	Experiment 2
Mean of A distribution*	46	45	45
Mean of B distribution*	54	55	55
Distal stimulus *SD**	5	0.01	0.01
Perception *SD*	0.0	0.0	0.0
Classification criterion *SD*	8.01	6.65	5.26
Confidence criteria *SD*	24.25	38.84	32.48
Equal-priors criterion	48.86	50.0	50.76
Criterion shift	±1.66	4.64	2.85
Confidence criteria	±{−13.73, −8.53, −4.09}	±{−66.7, −54.5, −29.6, −22.6, −0.76, 8.5}	±{−94.4, −39.2, −22.2, −22.2, −4.96, 27.9}

Note—The equal-priors criterion value is an absolute stimulus strength representing the default setting of the central criterion. The criterion shift reflects how much the central criterion moves in response to base rate manipulations, and the confidence criteria parameters represent the relative spacing of confidence criteria around the central classification criterion. *Represents fixed parameters determined by experimental conditions.

are several ways in which the model must be interpreted carefully. First, the model does not take into account response dependencies, which appear to be quite important in our task. Also, the model was fitted to the pooled data across 43 participants, so any individual differences could be misattributed to decision noise. Finally, the model assumes that the response policy is fixed relative to the central classification criterion, that the confidence criteria are all sampled with the same variance, and that they have normal distributions. None of these assumptions is likely to be true, although they are reasonable enough to produce accurate fits to data.

One benefit of model fitting is the ability to interpret the obtained parameter values. These fitted parameter values (shown in Table 1) indicate that the standard deviation of the confidence criteria was roughly three times as large as the standard deviation of the classification criterion, suggesting that confidence ratings are less reliable than stimulus classifications.

Given the ability of the model to account for our results, we also attempted to fit the model to previously published data sets that have been taken as evidence against SDT assumptions. We will show these fits next and use the goodness-of-fit and estimated parameters to evaluate whether the data present substantial problems for the notion of a flexible decision criterion.

Application of the DNM to Balakrishnan (1998a)

Our experiment produced most of the important violations of SDT that were noticed by Balakrishnan (1998a), and our modeling showed that the data could be produced by a model that did have a shifting response policy, if confidence noise was larger than classification noise. Our ROC analysis identified confidence-related processes as being the source of the violations, and our analysis of sequential dependencies also demonstrated an important response-related source of noise. However, our experiment did not exactly replicate earlier experiments by Balakrishnan (1998a, 1998b, 1999), since he used different types of discrimination tasks, adopted larger base rate manipulations, and instructed observers to avoid low-confidence responses. The primary consequence of these differences is that in our experiment, the observed changes in ROC functions were smaller (yet still reliable), and the upper bound for the largest detectable criterion shift was relatively large. Consequently, our experiment only had power to detect large changes. Thus, it is possible that the DNM might be unable to account for more challenging data, such as those reported by Balakrishnan (1998a).

We fitted the DNM to the C-ROC, $U_R(k)$, and probability correct functions from the two experiments reported by Balakrishnan (1998a). Without external noise, the scale of the perceptual distributions is somewhat arbitrary, so we used similar mean values as our previous fits in order to foster easier comparison. These observed and simulated functions are shown in the left column of Figure 12, and the obtained parameter values are shown in Table 1. This model used 11 fitted parameters to account for the response distribution across the 28 confidence ratings (14 in each base rate condition). Results showed an excellent fit to data, in particular showing the changes in β, but no change in $U_R(k)$ or suboptimal response probabilities. The model is also able to produce probability correct functions that fall below .5, which were not present in our experiment. The model slightly mispredicts the points on the high-probability confidence responses in the rare stimulus condition, indicating that our assumption that all confidence criteria shift equally relative to the classification criterion shift is probably incorrect. In addition, the observed C-ROC functions are somewhat more linear than the simulated functions, indicating that the human observers' confidence responses were less related to any internal evidence state than were those of the model. These two mispredictions are related: The model uses a large estimated confidence criterion variance and spacing between criteria to estimate the spacing between adjacent points on the C-ROC function; by doing so, it produces some discriminability between adjacent confidence ratings, and the resultant C-ROC function bows outward in response.

Balakrishnan (1998a) reported results from a second experiment showing similar violations of SDT. Again, the results (shown on the right side of Figure 12) are well

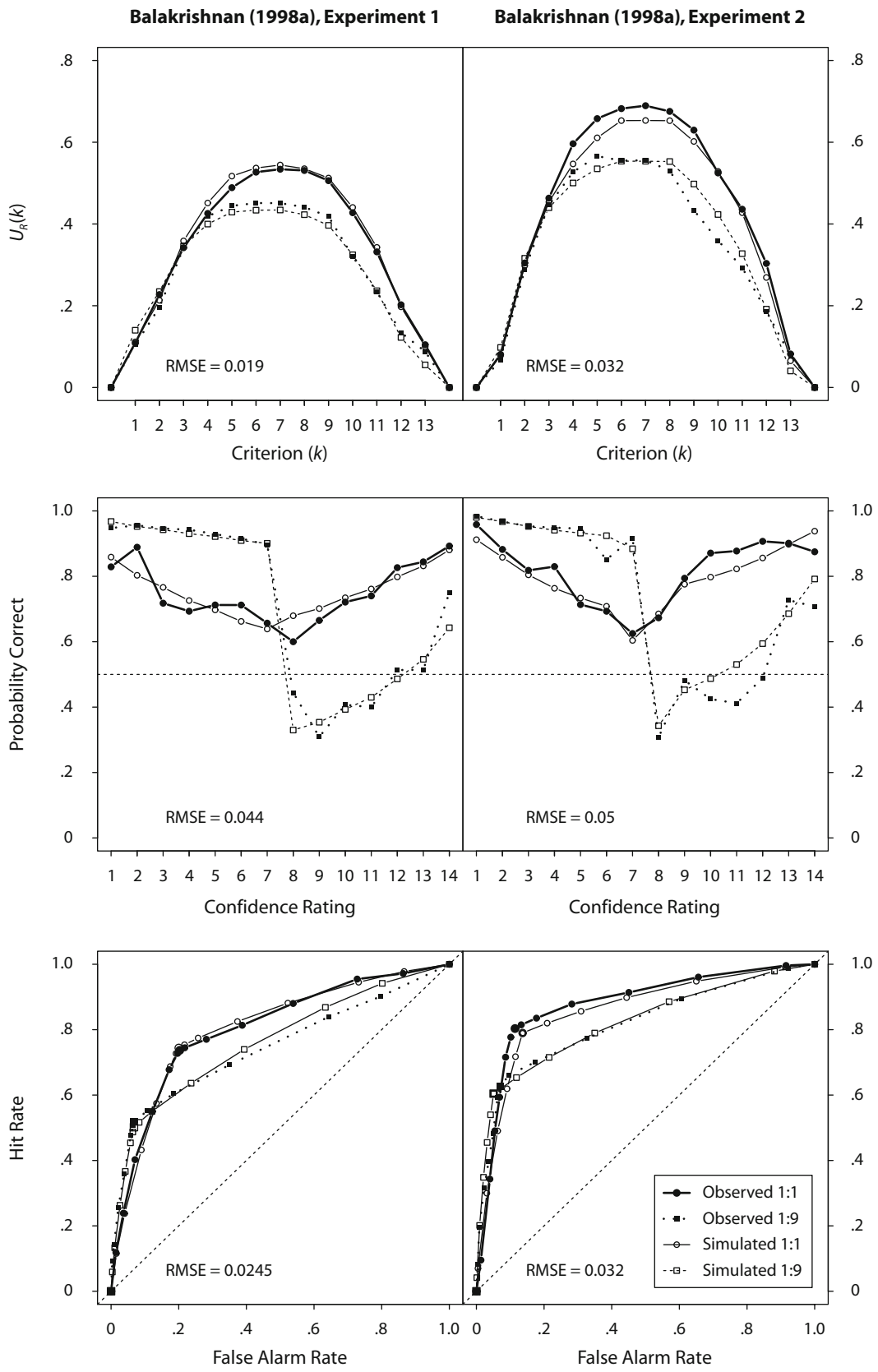

Figure 12. Decision noise model fitted to data from Experiments 1 and 2 of Balakrishnan (1998a), mixed successive condition. Left panels show the same data as those in Figure 1.

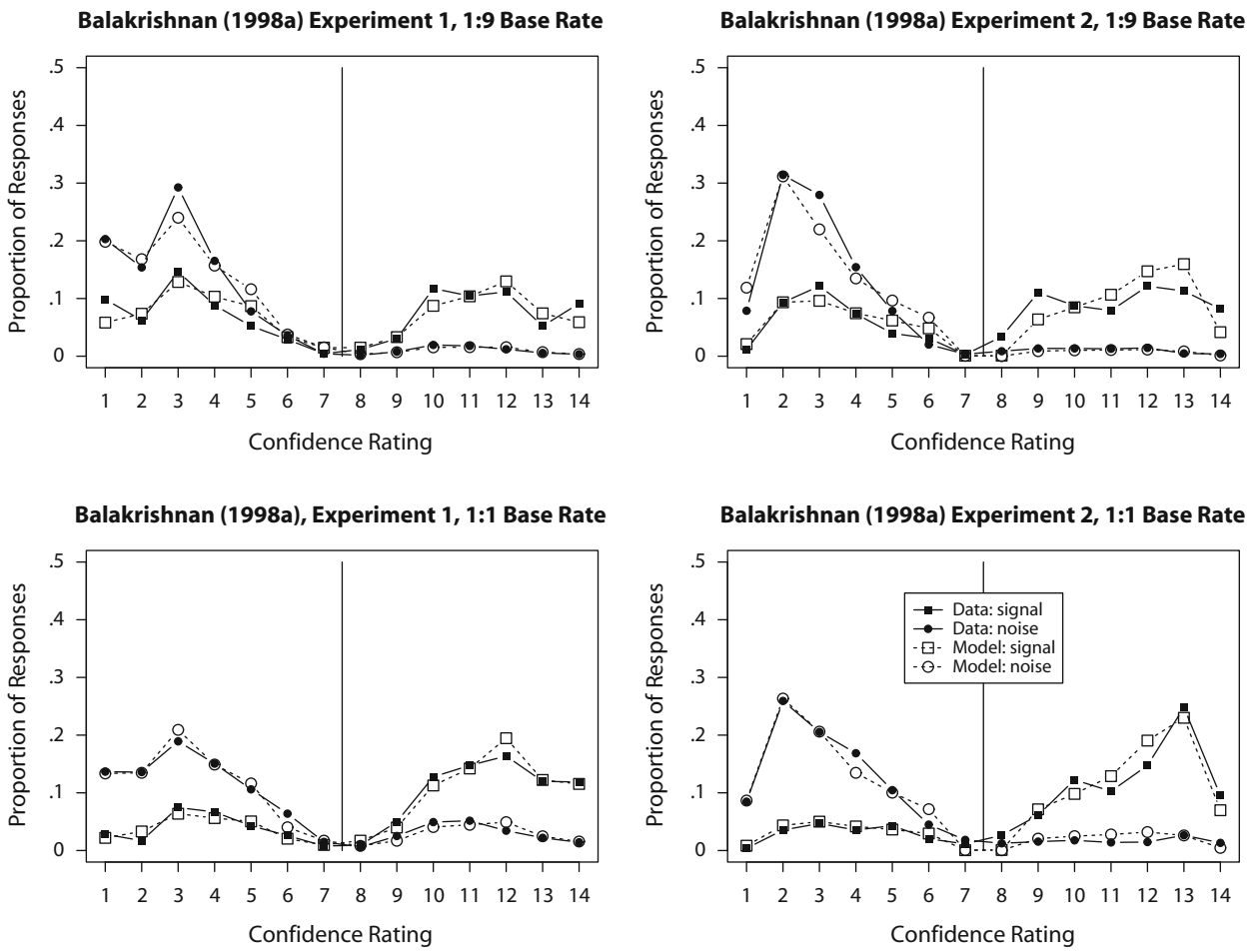

Figure 13. Decision noise model fitted to response proportions across confidence ratings from Experiments 1 and 2 of Balakrishnan (1998a), mixed successive condition.

accounted for by the DNM, and the fitted values were similar to those in the previous experiment, although the criterion shift was not as large.

In addition to fitting the statistics shown in Figure 12, it is important for us to show that the model can reproduce similar upper bound predictions about the amount of bias present in these experiments. In both of these experiments, not only was there little evidence for criterion shifts according to the $U_R(k)$ function, the experiments should have been powerful enough to detect shifts, because the upper bound on the proportion of biased responses was fairly small (i.e., around .015). This result is indexed by the overall proportion of least confident responses made in each experiment. Figure 13 shows the model's predictions about the probability distribution across responses for both conditions of both experiments. Predicted values for the least confident responses were small and similar to the observed values.

These demonstrations show that the DNM is able to account for the observed data across three experiments in fairly precise, quantitative ways. Thus, a version of SDT extended to incorporate decision noise appears to be a reasonably complete yet parsimonious account of the suboptimalities and biases found by Balakrishnan (1998a; see also Balakrishnan & MacDonald, 2002). Furthermore, in each case, the model accounted for the data by assuming that the criterion *did* shift, but that the shift was masked by decision noise.

Application of the DNM to Van Zandt (2000)

The previous model fits each examined the ability of the DNM to account for visual classification data pooled across participants for manipulations of base rate. However, so far we have ignored payoff manipulations, which have been shown to have effects similar to those of changes in base rate. Furthermore, pooling data across participants could hide individual variability that the model may be unable to capture. In order to demonstrate the power of the DNM for data from individuals, including payoff manipulations, we modeled the results of two experiments by Van Zandt (2000), who presented data from individual participants in a recognition memory experiment in which confidence ROC functions were collected and base rate or payoff structure was manipulated. Fits of the DNM to the data from 15 participants are shown in Figures 14–16.

Experiment 1 of Van Zandt (2000) included two between-participants conditions: fast and slow presenta-

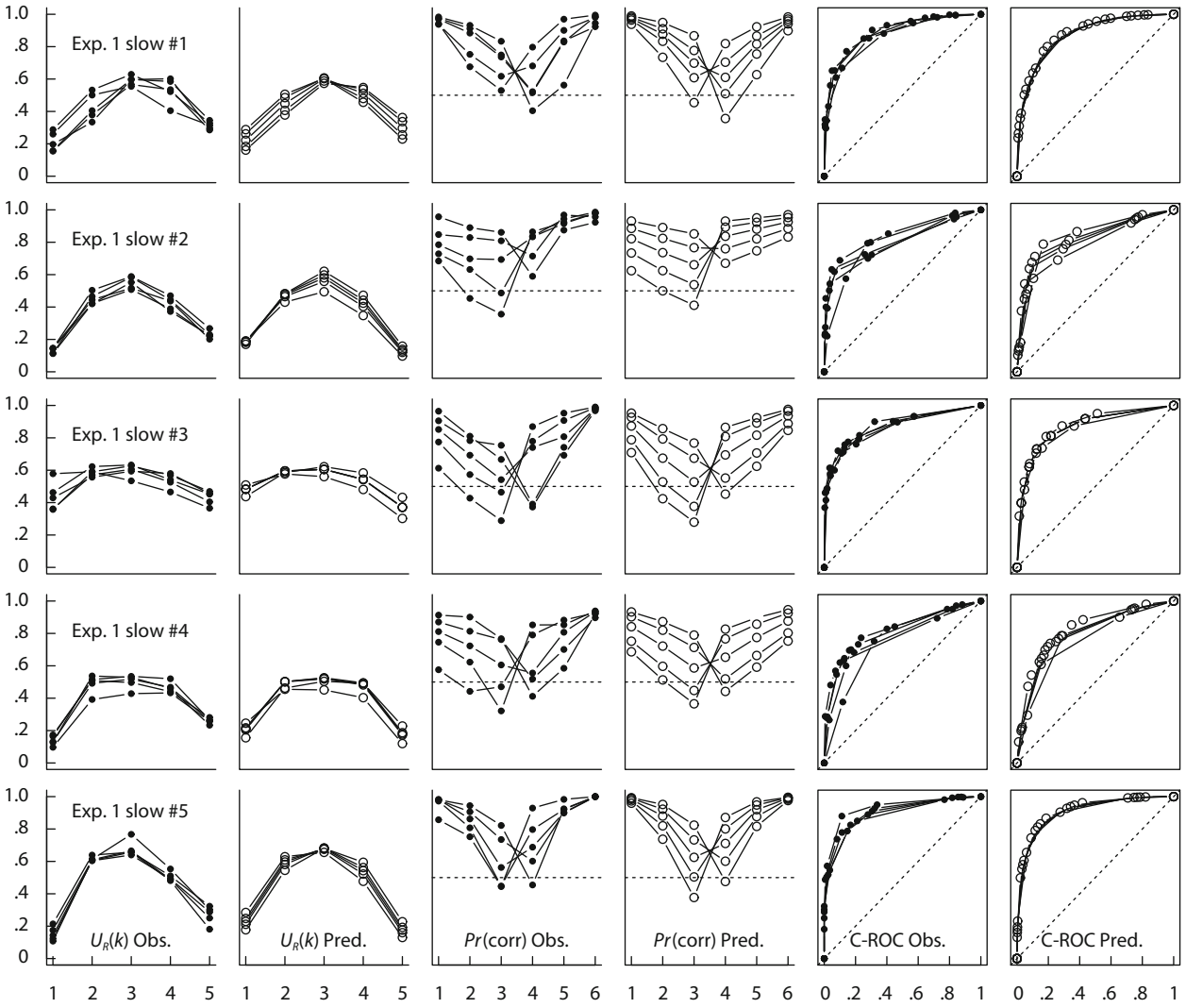

Figure 14. Decision noise model fitted to data from Experiment 1, slow condition of Van Zandt (2000). Each row depicts observed and fitted data from a different participant. The first two columns are observed and fitted $U_R(k)$ functions, columns 3 and 4 are observed and fitted probability correct, and columns 5 and 6 are observed and fitted C-ROC functions. Each connected sequence of points represents a different base rate condition.

tion of words. Five participants in each presentation condition were presented with twelve 32-word sets followed by 40-word test sets containing 8, 14, 20, 26, or 32 of the presented words. Another 5 participants took part in Experiment 2, which had five different within-participants payoff conditions (0, 1, 2, 3, or 4 points for hits, and 4, 3, 2, 1, or 0 points for correct rejections).

During the testing phase, participants were asked to rate on a 6-point Likert-type scale how certain they were that each word was old. Figures 14–16 show several statistics computed on these data across the different conditions and participants, as well as the DNM fit to the data.

Overall, the DNM provides an adequate fit to the data. Notably, substantial individual variability exists across the participants, and the model is able to capture this variability with corresponding changes in parameter values. Although it is difficult to see, substantial crossover effects in the C-ROC functions are produced both by several human participants and by the DNM.

The values of the fitted parameters for the DNM are useful for interpreting the results. For these data, we estimated two decision noise parameters: one for the central classification criterion, and a second for all the remaining confidence criteria. Additionally, we estimated three location parameters: one for the central equal-odds/equal-base rate criterion, and two for the means of the four confidence criteria symmetrically distributed around the center. Thus, the baseline response policy was completely specified by five parameters. Additionally, the five condition manipulations (base rate or payoff) were fit by assuming that the entire response policy shifted its mean symmetrically about the equal-odds or equal-payoff point for the different conditions, resulting in two additional parameters. Perceptual noise variance was assumed to be 0 for these simula-

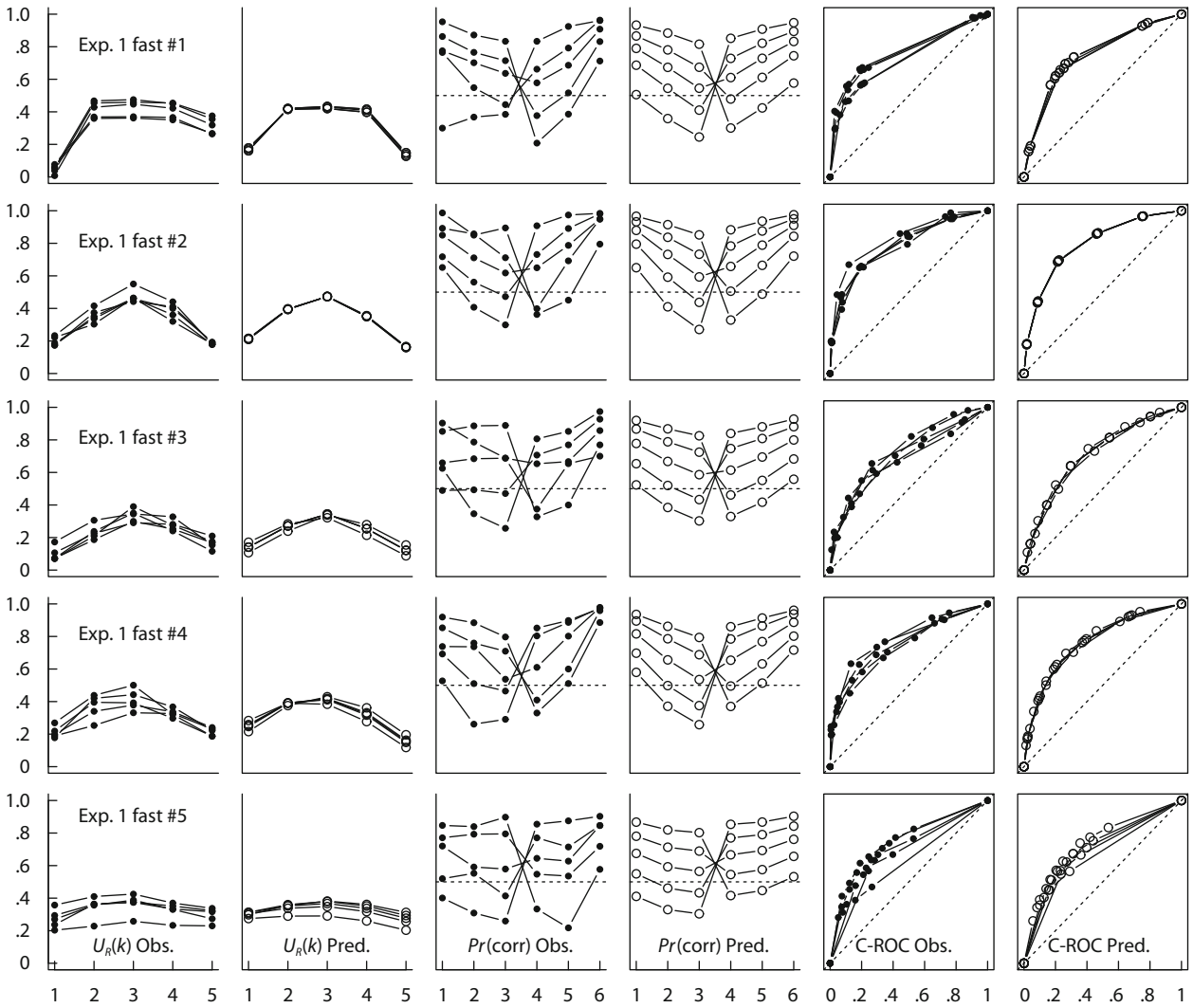

Figure 15. Decision noise model fitted to data from Experiment 1, fast condition of Van Zandt (2000). Each row depicts observed and fitted data from a different participant. The first two columns are observed and fitted $U_R(k)$ functions, columns 3 and 4 are observed and fitted probability correct, and columns 5 and 6 are observed and fitted C-ROC functions. Each connected sequence of points represents a different base rate condition.

tions. Thus, a total of seven free parameters were estimated for each participant, which accounted for the 50 degrees of freedom of each C-ROC function in Figures 14–16. Five additional degrees of freedom in Experiment 1 and one degree of freedom in Experiment 2 are present in the probability correct data; thus, 51–55 degrees of freedom in the data were fitted for each participant using seven free parameters. Table 2 shows the fitted parameter values for all participants. Of special interest is the ratio between the classification noise and the confidence noise. As in the previous experiments, for each subject in each condition, this ratio was greater than 1.0—ranging from 1.37 to 10.89—with a mean of 3.58, suggesting that confidence noise is greater than classification noise. Furthermore, differences between conditions were attributed to shifts in the decision criterion, even though few of the $U_R(k)$ functions suggest that the criterion shifted.

Thus, our application of the DNM to the data in Van Zandt (2000) demonstrates that the effects revealed by the model hold up at an individual-participant level and are not a result of averaging across participants. Indeed, the model fit individual participants well with relatively few free parameters. Additionally, the model is able to account for violations of SDT stemming from payoff manipulations, as well as base rate manipulations, in a recognition memory task.

DISCUSSION

SDT has become—and is likely to remain—an important theoretical framework and a widely adopted data analysis tool, despite its many weaknesses. Its greatest strength and greatest weakness is its simplicity: It provides a parsimonious and easily applied account of perception and memory

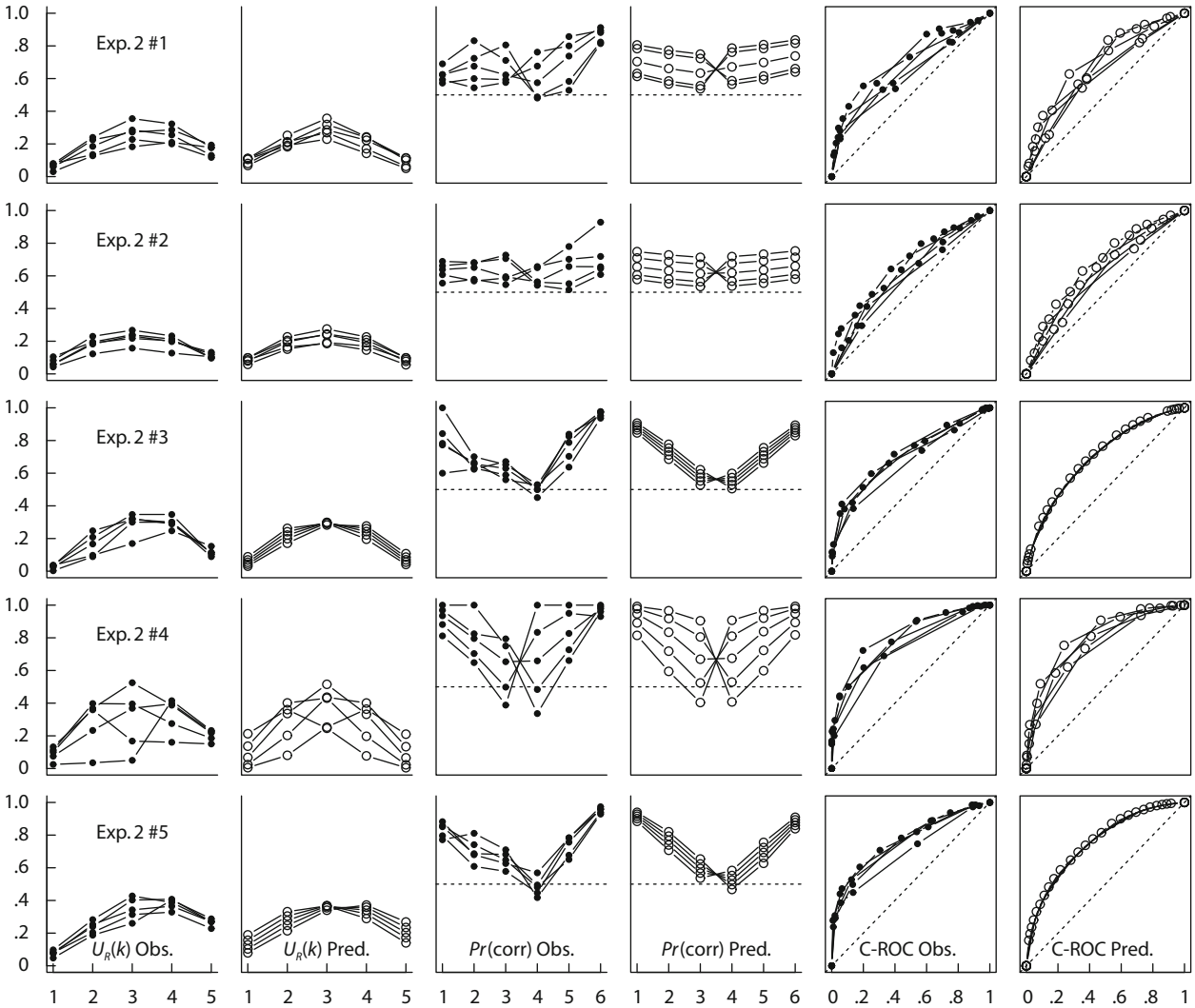

Figure 16. Decision noise model fitted to data from Experiment 2 of Van Zandt (2000). Each row depicts observed and fitted data from a different participant. The first two columns are observed and fitted $U_R(k)$ functions, columns 3 and 4 are observed and fitted probability correct, and columns 5 and 6 are observed and fitted C-ROC functions. Each connected sequence of points represents a different payoff condition.

retrieval that captures almost none of the true detail of the underlying processes. Balakrishnan's (1998a, 1998b, 1999) extensive research on the assumptions and violations of SDT has been an important contribution to the literature on SDT and should not be cast aside lightly. Indeed, it has shown that there are profound suboptimalities and biases in true responding that are not captured by SDT. Yet, many of his critiques are made on the basis of data from confidence rating procedures, which add an additional layer of complexity (and an additional set of assumptions) to their interpretation. Thus, it is possible that the apparent violations of the assumptions made by SDT are really violations of the assumptions about how confidence judgments are made. Insofar as classic SDT assumes that responses are under strategic control, it is reasonable for one to incorporate suboptimal response policies into an extension of the theory, and, by doing so, we have shown that many of the apparent violations of SDT can be explained.

Our research has uncovered three important results suggesting that there is substantial uncertainty in response processes: (1) The observed DS-ROC functions do not change in response to base rate manipulations, despite changes in C-ROC functions; (2) a high proportion of sequential dependencies were found in confidence responses; and (3) the DNM, which extended SDT to confidence ratings with decision noise, accounted for these data patterns across multiple data sets by assuming decision-related noise. Together, these results suggest that participants do not use static evidence criteria to make responses, and they are especially unable or unwilling to use confidence ratings in a consistent way. Instead, great uncertainty exists between different confidence states, and confidence responses appear to depend on factors other than the internal level of perceptual evidence. Next, we will examine each of our main findings in greater detail.

Table 2
Parameter Estimates of the Decision Noise Model to the Data From Van Zandt (2000)

Experiment	Participant	Classification Noise	Confidence Noise	Ratio	±Med. Conf. Criteria	±High Conf. Criteria	Equal-Odds Criteria	±Med. Odds Criteria	±Extreme Criteria
1 slow	1	5.852	10.149	1.73	0.451	5.123	49.195	−0.872	−1.431
1 slow	2	5.668	33.106	5.84	−15.965	20.622	52.423	−0.591	−1.962
1 slow	3	5.674	13.471	2.37	−10.426	−1.399	51.509	−0.018	−1.447
1 slow	4	6.959	23.425	3.37	−25.762	14.718	50.976	−0.421	−3.271
1 slow	5	4.991	8.758	1.75	−0.538	8.846	50.584	−0.379	−1.125
1 fast	1	8.717	28.078	3.22	−35.239	19.351	51.440	−0.850	−0.919
1 fast	2	7.827	16.653	2.13	−1.876	7.710	51.092	0.094	−0.134
1 fast	3	11.182	29.165	2.61	−5.241	11.032	50.983	−0.020	−3.378
1 fast	4	8.796	19.088	2.17	−6.606	6.098	52.468	0.336	−1.982
1 fast	5	9.939	49.093	4.94	−59.961	−37.480	53.216	−1.429	−4.559
2	1	10.700	82.676	7.73	−27.602	28.004	51.501	−6.948	−9.032
2	2	14.227	154.924	10.89	−125.264	43.671	50.289	−7.341	−12.567
2	3	13.029	19.216	1.47	3.121	20.635	49.134	−1.640	−3.575
2	4	7.155	15.427	2.16	3.757	19.437	50.091	−4.567	−9.337
2	5	10.425	14.257	1.37	0.998	7.011	48.365	−1.294	−2.705

External Noise and Distal-Stimulus-Related Statistics

We showed that the source of the ROC asymmetry can be identified by comparing DS-ROC and C-ROC functions, and we found that for our data, these asymmetries stemmed from response processes. Statistics related to the DS-ROC function have seen little use in SDT analysis, although they are easy to compute as long as external noise is added to the stimuli and the level of this noise on each trial is recorded. Although we did not do so in the present article, one could use the area under this function as a measure of perceptual-response sensitivity akin to A', A (Zhang & Mueller, 2005), or d'. If our conclusions are correct and decision noise plays a substantial role in signal detection tasks, then such stimulus-related statistics offer an important tool for dissociating changes in perceptual and decision noise across conditions.

Sensitivity measures depend critically on what factor is used as a measure of the distal stimulus, because they indicate the extent to which this variable is related to the information used to make the response. Although we used a value that was maximally diagnostic for our task, related functions could be computed for other variables, and the area under the function can serve as an index of the extent to which a variable affects discriminability. This could be especially useful in memory tasks, where DS-ROC functions could be computed for different psycholinguistic measures (e.g., frequency, imageability) or other manipulations (e.g., study time, repetitions) and could index the extent to which a specific variable is related to response discriminability. Variables completely unrelated to discrimination produce an area close to 0.5, whereas variables more closely related to discrimination produce areas closer to 1.0.

Sequential Dependencies

We also found substantial sequential dependencies in the responses made during our task. Such dependencies are common in many perceptual tasks and could have arisen for a number of reasons. One possibility is that these dependencies are fundamental normative aspects of human responding that have been learned through experience or that have evolved over the course of phylogeny because they are useful in many important situations. This may include shifts in classification criteria that occur as a result of learning and feedback. Alternately, they may reflect optimal strategic responses to improperly perceived sequential dependencies. A third explanation is that they may arise because of a need to maintain consistency: Two consecutive stimuli that are perceptually similar may be given a similar confidence rating. Participants may in fact not be aware of the perceptual evidence used to make their decisions and may have essentially only two discriminable states that confidence ratings can be produced from (an argument that originated during the debates between threshold theory and SDT; see Malmberg, 2002, for a review). Thus, the sequential dependencies may not be a fundamental property of the perceptual or response processes, but a side effect of the task that asks participants to distinguish between indiscriminable states.

The DNM

In this article, we presented a new extension of SDT that we call the DNM, which allowed different sources of noise to be parametrically manipulated and estimated from data. We showed that this model could fit observed data, and suggested that confidence-related noise may be the source of the crossover ROC functions and the observed failure of $U_R(k)$ peaks to move in response to base rate or payoff manipulations.

Most importantly, the model revealed that across a range of participants, experimental paradigms, and researchers, noise related to confidence assessment was consistently larger than noise related to two-category classification. This result indicates that confidence ratings are relatively unreliable and may lack power for testing assumptions about underlying perceptual distributions. Consequently, the model shows that decision noise is a factor that should be taken seriously in order to understand decision making under uncertainty.

Response Time Measures and Sequential Sampling Models

The models and statistics we have discussed in this article have not dealt directly with RT, which is an important factor in decision theory. There are two broad areas in which RT analysis may provide insights into the issues covered presently: First, sequential sampling models that make predictions about RTs may provide important insights into the phenomena we have described in this article; second, RT measures may provide a convenient alternative to confidence ratings.

Van Zandt (2000), for example, accounted for her data with a sequential sampling model. In general, these models differ from SDT in their assumption that the perceptual state changes over time as evidence about a stimulus is sampled. In comparison with SDT, in which base rate and payoff are assumed to only influence the placement of the decision criterion, sequential sampling models offer considerably more flexibility in how base rates and payoffs can be modeled. Diederich and Busemeyer (2006), for example, compared three different ways in which a sequential sampling framework could account for the effects of base rate manipulations in a perceptual task. However, just like SDT itself, such sampling models need to be augmented with additional assumptions in order to produce confidence ratings. Regardless, we have shown that decision noise plays an important role in signal detection tasks and can account for the violations of SDT theory observed by Balakrishnan (1998a, 1998b, 1999). Such noise can be implemented in models that assume sequential evidence sampling, and therefore our conclusions are not limited to SDT.

One way in which a sequential sampling model might produce confidence ratings is as a function of the time taken for accumulated evidence to reach a decision criterion. Consequently, it is possible to use RT as a proxy for confidence ratings or strength of the percept, and to form stimulus-related ROC functions using the techniques we discussed earlier. Although RT to a binary decision might not be subject to the same decision noise we observed for confidence judgments, it is unrealistic for one to assume that the mapping between the strength of the percept or confidence and the associated RT is noise free. Thus, the issues associated with testing the assumptions of SDT are not unique to confidence ratings.

Summary and Conclusions

Our findings cast a new light on some recent criticisms of SDT (Balakrishnan, 1998a, 1998b, 1999). We showed that these violations are attributable to decision noise, which suggests that confidence ratings are not appropriate for forming ROC functions. Yet, our results did not unambiguously show that the underlying assumptions of SDT were correct; they simply show that previous attempts to test them have been inadequate and, consequently, that the model is more difficult to test than previously thought. Furthermore, the model we propose that does account for the results is not SDT itself; rather, it is an extension of SDT incorporating decision noise—a component that many previous researchers have criticized SDT for ignoring (a similar argument was made recently by Benjamin, Diaz, & Wee, 2007).

We believe that the time has come to acknowledge the importance of decision noise in signal detection tasks and to begin moving beyond the simple application of SDT. Even if the basic assumptions of SDT are correct, whenever decision noise exists, d' and β may be unable to separate perceptual and decision processes in any meaningful way, and will thus be of little use. Balakrishnan's (1998a, 1998b, 1999) proposed statistics were intended to test for the existence of suboptimality in the decision rule, and are thus unable to distinguish between perceptual and decision noise as well. We hope that through techniques using external noise (like the DS-ROC function used in the present article), future research may be better able to isolate influences of the perceptual and response systems and to arrive at a clearer understanding of human perception and choice under uncertainty.

AUTHOR NOTE

S.T.M. is now employed at Klein Associates Division of ARA Inc. (Fairborn, OH). Much of the research presented in the present article was carried out while both authors were affiliated with the Department of Psychological and Brain Sciences at Indiana University, Bloomington. This research was supported by NIMH Grant MH12717, and by a postdoctoral fellowship to C.T.W. from the German Academic Exchange Service (DAAD). The authors thank Richard Shiffrin, Kelly Addis, Jerry Balakrishnan, William Estes, Martin Gevonden, Krystal Klein, Angela Nelson, Adam Sanborn, and Erin Shubel for helpful comments and discussions, as well as Jerry Balakrishnan and Trish Van Zandt for making their data available. Address correspondence to S. T. Mueller, Klein Associates Division, ARA Inc., 1750 N. Commerce Center Blvd., Fairborn, OH 45324 (e-mail: smueller@ara.com).

REFERENCES

BALAKRISHNAN, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, **40**, 601-623.

BALAKRISHNAN, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, **3**, 68-90.

BALAKRISHNAN, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 1189-1206.

BALAKRISHNAN, J. D., & MACDONALD, J. A. (2002). Decision criteria do not shift: Reply to Treisman. *Psychonomic Bulletin & Review*, **9**, 858-865.

BENJAMIN, A. S., DIAZ, M., & WEE, S. (2007). *Signal detection with criterion variability: Applications to recognition*. Manuscript submitted for publication.

BUSEMEYER, J. R., & MYUNG, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, **121**, 177-194.

DIEDERICH, A., & BUSEMEYER, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics*, **68**, 194-207.

DORFMAN, D. D., & BIDERMAN, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology*, **8**, 264-284.

DORFMAN, D. D., SASLOW, C. F., & SIMPSON, J. C. (1975). Learning models for a continuum of sensory states reexamined. *Journal of Mathematical Psychology*, **12**, 178-211.

DUSOIR, A. E. (1974). Thomas and Legge's matching hypothesis for de-

tection and recognition tasks: Two tests. *Perception & Psychophysics*, **16**, 466-470.
EREV, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review*, **105**, 280-298.
GOLD, J. M., MURRAY, R. F., BENNETT, P. J., & SEKULER, A. B. (2000). Deriving behavioral receptive fields for visually completed contours. *Current Biology*, **10**, 663-666.
GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
HEALY, A. F., & KUBOVY, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 344-354.
JONES, M., LOVE, B. C., & MADDOX, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 316-332.
KAC, M. (1969). Some mathematical models in science. *Science*, **166**, 695-699.
KUBOVY, M., & HEALY, A. F. (1977). The decision rule in probabilistic categorization: What it is and how it is learned. *Journal of Experimental Psychology: General*, **106**, 427-446.
LARKIN, W. (1971). Response mechanisms in detection experiments. *Journal of Experimental Psychology*, **91**, 140-153.
LEE, W. (1963). Choosing among confusably distributed stimuli with specified likelihood ratios. *Perceptual & Motor Skills*, **16**, 445-467.
LEE, W., & JANKE, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology*, **68**, 376-382.
LEE, W., & ZENTALL, T. R. (1966). Factorial effects in the categorization of externally distributed stimulus samples. *Perception & Psychophysics*, **1**, 120-124.
LOH, S., LAMOND, N., DORIAN, J., ROACH, G., & DAWSON, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers*, **36**, 339-346.
LU, Z. L., & DOSHER, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, **38**, 1183-1198.
MALMBERG, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 380-387.
MUELLER, S. T. (1998). *Reinforcement learning in signal detection*. Unpublished master's thesis, University of Michigan, Ann Arbor.
MUELLER, S. T. (2005). PEBL: The psychology experiment building language (Version 0.05) [Computer experiment programming language]. Retrieved August 2005 from pebl.sourceforge.net.
NEWSOME, W. T., BRITTEN, K. H., & MOVSHON, J. A. (1989). Neural correlates of a perceptual decision. *Nature*, **341**, 52-54.
PARDUCCI, A., & SANDUSKY, A. (1965). Distribution and sequence effects in judgment. *Journal of Experimental Psychology*, **69**, 450-459.
PHILIASTIDES, M. G., & SAJDA, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, **16**, 509-518.
R DEVELOPMENT CORE TEAM (2006). R: A language and environment for statistical computing (Reference index version 2.6.2). Vienna: R Foundation for Statistical Computing.
SANDUSKY, A. (1971). Signal recognition models compared for random and Markov presentation sequences. *Perception & Psychophysics*, **10**, 339-347.
SCHOEFFLER, M. S. (1965). Theory for psychophysical learning. *Journal of the Acoustical Society of America*, **37**, 1124-1133.
THOMAS, E. A. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology*, **10**, 241-264.
THOMAS, E. A. (1975). Criterion adjustment and probability matching. *Perception & Psychophysics*, **18**, 158-162.
TREISMAN, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review*, **9**, 845-857.
TREISMAN, M., & WILLIAMS, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, **91**, 68-111.
VAN ZANDT, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 582-600.
WARD, L. M. (1973). Use of Markov-encoded sequential information in numerical signal detection. *Perception & Psychophysics*, **14**, 337-342.
ZHANG, J., & MUELLER, S. T. (2005). A note on ROC analysis and nonparametric estimate of sensitivity. *Psychometrika*, **70**, 145-154.
ZHANG, J., RIEHLE, A., & REQUIN, J. (1997). Analyzing neuronal processing locus in stimulus–response association tasks. *Journal of Mathematical Psychology*, **41**, 219-236.

NOTES

1. In this article, we use the term *signal detection task* in a loose sense, referring to the basic paradigm in which two stimulus classes are discriminated and categorized into two classes (old or new, signal or noise, yes or no, A or B, etc.). These can include classical detection tasks (discriminating signal presence from absence), a wide range of two-alternative forced choice tasks, recognition memory tasks, and other stimulus classification tasks. Among other similarities, these tasks share the property that a version of SDT is often used as a theoretical account of the resulting data.

2. The peak observed in C-ROC functions normally corresponds to the point at which the slope of the function crosses from being greater than 1 to being less than 1; thus, it is an important equal-likelihood point on the ROC function.

3. The fact that the average of multiple points on a downwardly convex function like a ROC function must lie below the function is known as *Jensen's inequality*. Jensen's inequality implies that decision noise will produce an observed ROC function that lies below some ideal latent ROC function formed by the perceptual evidence alone and will approach this latent function to the extent that a static decision rule is followed.

4. The basic steps of a proof showing the correspondence between the C-ROC function and $U_R(k)$ are as follows: For $0 < x < y < 1$, the distance between the point (x,y) and the line $y = x$ along the line with slope -1 is

$$\sqrt{1/2}(y-x).$$

Thus, any point on an ROC function has a distance to the diagonal proportional to the difference between x and y. The C-ROC function plots cumulative distributions for signal and noise against one another at each confidence rating, whereas $U_R(k)$ is formed by the difference between these same cumulative distributions at each confidence rating. Consequently, the distance from an ROC point to the diagonal is proportional to its corresponding $U_R(k)$ value.

5. Because of the normal distribution of the confidence criteria, the sampled criteria are not constrained to be ordered according to the confidence level they represent. If a criterion for more extreme confidence level is closer to the neutral point than is the previously sampled criterion, then it is also necessarily smaller than the percept, and the next criterion level is sampled. This procedure was chosen solely for simplicity and convenience, and details are given in Appendix A.

6. When interpreting single-condition ROC functions on the basis of confidence, researchers often implicitly assume that a confidence rating made after the stimulus is presented is equivalent to a criterion shift made before the stimulus appears. If these are not the same, then confidence-based ROC functions may not correspond to ROC functions produced by manipulating base rate or payoff.

7. The degrees of freedom of the C-ROC analysis were 82 instead of 84, because the z-ROC slope from one condition of 1 participant was undefined (only two different confidence ratings were used by this participant in this condition).

8. Recently, Benjamin, Diaz, and Wee (2007) noted that reasonable models of signal detection can be formed using decision noise instead of perceptual noise. Although decision noise models are difficult to distinguish from perceptual noise models, Benjamin et al. ultimately concluded that decision-noise-only models were untenable.

(*Continued on next page*)

APPENDIX A
The Decision Noise Model

The decision noise model is a simple attempt to add confidence judgments to a signal detection model while allowing for environmentally introduced external noise, perceptual noise, and decision noise. Many different methods, distributions, and evidence combination rules could be used to simulate these processes, and the differences between them are beyond the scope of this article. Our model generalizes these processes, using a simple account of criterion variability. To begin, we define the signal and noise distributions (i.e., the distal stimuli) with the possibility of externally introduced noise. In signal detection theory (SDT), it is typically assumed that this noise is 0 or that it is confounded with perceptual noise, but in practice, signal detection tasks often introduce such noise in a measurable way. For the density functions (f) shown below, μ and σ refer to mean and standard deviation of the normal distribution—$N(\mu, \sigma)$. If stimuli are presented without external noise, then the value of σ would be 0.

$$f_{\text{distal signal stimulus}} = N(\mu_S, \sigma_S)$$

$$f_{\text{distal noise stimulus}} = N(\mu_N, \sigma_N)$$

Next, we define the perceptual noise distribution and compute the perceptual stimulus distributions. In classic SDT, the perceptual noise distribution is typically assumed to be the primary factor influencing the d' statistic. With $\otimes$ denoting the convolution operation between two density functions, the perceptual stimulus distributions produced by successive random sampling from external noise and perceptual noise are defined in Equations A1–A3.

$$f_{\text{perceptual noise}} = N(0, \sigma_{\text{perc}}) \tag{A1}$$

$$f_{\text{signal perceptual distribution}} = f_{\text{distal signal stimulus}} \otimes f_{\text{perceptual noise}} \tag{A2}$$

$$f_{\text{noise perceptual distribution}} = f_{\text{distal noise stimulus}} \otimes f_{\text{perceptual noise}} \tag{A3}$$

In Equation A1, σ_{perc} refers to the standard deviation of the perceptual noise. Now, a response policy must be defined that describes the probability of choosing each response given a perceptual stimulus intensity. For a two-alternative classification, we assume that the response is made by comparison to a classification criterion that is sampled from a normal distribution whose mean and variance are free parameters. Here, we define

$$F(x) = \int_{-\infty}^{x} f(y)\,dy.$$

The density of the sampling distribution is defined in Equation A4, and the resultant response policies are defined in Equations A5 and A6.

$$f_{\text{classification criterion}} = N(\mu_{\text{class}} + \delta_{\text{condition}}, \sigma_{\text{class}}) \tag{A4}$$

$$p_{\text{yes}}(x) = F_{\text{classification criterion}}(x) \tag{A5}$$

$$p_{\text{no}}(x) = 1 - p_{\text{yes}}(x) \tag{A6}$$

In Equation A4, μ_{class} and σ_{class} refer to the baseline mean and standard deviation of the classification criterion, and $\delta_{\text{condition}}$ refers to the shift in the mean associated with the respective experimental condition. $p_{\text{yes}}(x)$ indicates the probability of giving a "yes" response if the perceptual stimulus has value x. Next, we assume that confidence responses are controlled by comparison to multiple sampled confidence criteria through a conditional comparison process. We assume that there are $2K$ confidence criteria with distributions whose means are distributed symmetrically around the classification criterion (and which are free parameters in the model). After the classification criterion is compared with the percept, the locations of confidence criteria are sampled, and consecutive comparisons are made with successive criteria until a criterion exceeds the sampled percept (by which we mean that it is further away from the sampled response criterion than is the percept). The criteria are examined in nominal order from the center outward; their actual sampled values may fall in another order and, indeed, their means are not restricted to follow any specific order. As a notational convenience, we define the absolute decrement function $\text{dec}(i) = \text{sign}(i)[\text{abs}(i) - 1]$ for $\text{abs}(i) \geq 1$, so that if i is 4, then $\text{dec}(i)$ is 3, and if i is -3, then $\text{dec}(i)$ is -2. Then, the conditional densities of the sampled criteria are shown in Equations A7 and A8.

$$f_{k_0} = f_{\text{classification criterion}} \tag{A7}$$

$$f_{k_i | x \text{ exceeds } k_{\text{dec}(i)}} = N(\mu_{\text{class}} + \delta_{\text{condition}} + \delta_{k_i}, \sigma_{\text{conf}}) \tag{A8}$$

$$i \in \{-K, \ldots, -2, -1, 1, 2, \ldots, K\}$$

$$\delta_{k_i} = -\delta_{k_{-i}}$$

In Equation A8, δ_{k_i} refers to the difference between the mean of the classification criterion and the mean of confidence criterion k_i, and σ_{conf} refers to the standard deviation of the confidence criteria (i.e., confidence noise).

APPENDIX A (Continued)

The unconditional probability of a sampled percept exceeding each criterion can be calculated as:

$$F_{k_i}(x) = F_{k_i|x \text{ exceeds } k_{\text{dec}(i)}}(x) \times F_{x \text{ exceeds } k_{\text{dec}(i)}}(x), \tag{A9}$$

where

$$F_{k_i}(x) = \begin{cases} F_{\text{classification criterion}}(x); & i = 0, \\ F_{k_i|x>k_{i-1}}(x) \times F_{k_{i-1}}(x); & i > 0, \\ 1 - \left(1 - F_{k_i|x<k_{i+1}}(x)\right) \times \left(1 - F_{k_{i+1}}(x)\right); & i < 0. \end{cases}$$

Finally, in order to compute the probability of each confidence response being made given a percept x, we compute for each $i \in \{-K, \ldots, -2, -1, 1, 2, \ldots, K\}$:

$$p_{\text{confidence}_{-(K+1)}}(x) = 1 - F_{k_{-K}}(x)$$

$$p_{\text{confidence}_i}(x) = F_{k_i}(x) - F_{k_{\text{dec}(i)}}(x)$$

$$p_{\text{confidence}_{(K+1)}}(x) = F_{k_K}(x).$$

Simulated response policy functions computed with this formula are plotted in the main text in Figure 2. A central classification criterion and $2K$ confidence criteria generate $2(K + 1)$ confidence responses. Note that there is no equal-confidence response in this model, but one could be added by ignoring the medial classification criterion and treating the percepts between the two innermost sampled criteria as the equivocal region.

It is easily verified that for every x, $\Sigma_i p_{\text{confidence}_i}(x) = 1$, because all values of $F_{k_i}(x)$ are between 0 and 1, so the progression forms a telescoping series in which all intermediate terms cancel. Thus, for any x, $p_{\text{confidence}_i}(x)$ is a probability distribution over i. Together, $p_{\text{confidence}_i}(x)$ define a complete response policy, which can be combined with signal and noise perceptual distributions and stimulus base rates to estimate signal detection parameters either through a Monte Carlo simulation or numerical integration. For example, the C-ROC functions are defined by the ordered pairs

$$\left(\sum_{j=i}^{N} p_{\text{noise}}(j), \sum_{j=i}^{N} p_{\text{signal}}(j) \right)$$

for each criterion i of the N confidence criteria, where

$$p_{\text{signal}}(i) = \int_x p_{\text{confidence}_i}(x) \times f_{\text{signal perceptual distribution}}(x)\, dx$$

and

$$p_{\text{noise}}(i) = \int_x p_{\text{confidence}_i}(x) \times f_{\text{noise perceptual distribution}}(x)\, dx.$$

The functions $U_R(k)$ can be computed directly from the C-ROC function by computing

$$\sum_{j=k}^{N} p_{\text{signal}}(j) - \sum_{j=k}^{N} p_{\text{noise}}(j).$$

As defined, the decision noise model can be simulated via Monte Carlo techniques so that individual trials are simulated by sampling perceptual, external, and decision noise. This process is extremely computationally intensive; thus, instead, we implemented the models in the statistical computing language R (R Development Core Team, 2006) by numerically computing the densities of the distributions and by computing the effect of consecutive random samples on these densities using the `convolve()` function in R. This function computes a numerical convolution of two distribution densities using a fast Fourier transform operator. Doing so allows us to quickly obtain highly accurate estimates of the probability distributions for each confidence response, enabling data fitting through Newton-like methods using the R `nlm` function. Because perceptual noise was set to 0 for the simulations presented here, the convolution operation is actually unnecessary, since the distal stimulus is identical to the percept.

(*Continued on next page*)

APPENDIX B
Procedures for Computing ROC Functions

If data from an experiment having multiple confidence ratings and multiple external noise levels are cross-tabulated, then the DS-ROC function is computed by applying the same steps used to form the C-ROC function, but on the rows instead of the columns (or vice versa). In order to illustrate this correspondence, we will first show how C-ROC functions are formed on a sample data set, and then how DS-ROC functions are formed from the same data set.

Suppose that data were collected in an experiment in which visual stimulus intensity had six distinct values (100, 200, 300, 400, 500, or 600) and in which responses were made on a 6-point confidence scale, with 1–3 indicating a "low" response, and 4–6 indicating a "high" response. Table B1 shows hypothetical results from such an experiment with 200 trials.

Table B1
Cross-Tabulated Data From a Hypothetical Experiment With Six Confidence States (1–6) and Six Distal Stimulus Intensities (100–600)

	Rating					
	Low			High		
Intensity	1	2	3	4	5	6
100	5	5	6	2	3	2
200	9	15	8	3	6	3
300	4	8	13	2	3	3
400	2	3	6	9	7	5
500	2	6	2	9	11	6
600	1	3	2	7	7	12

Note—Each cell represents the number of trials that had a specific stimulus intensity and confidence response.

Computing the C-ROC Function

The C-ROC is formed by first dividing the trials into two classes, on the basis of either stimulus intensity or some nominal class that the actual stimulus was sampled from. In our example, we classify stimuli with an intensity of 300 or less as "noise," and stimuli with higher intensities as "signal." These two stimulus classes are shown in Table B2.

Table B2
Distribution of Confidence Ratings for Noise and Signal Trials

	Rating						
	Low			High			
Trial Type	1	2	3	4	5	6	Σ
Noise	18	28	27	7	12	8	100
Signal	5	12	10	25	25	23	100

Hit rate and false alarm rates can be computed from these distributions once a decision criterion is selected. For example, if the criterion were placed between confidence levels 2 and 3, then the hit rate (signal trials correctly called signal) would be 83/100. Likewise, the false alarm rate (noise trials incorrectly called signals) would be 54/100. In this analysis, we examine all hypothetical decision criteria, which consist of the points that divide adjacent confidence states. Table B3 shows hit rate and false alarm rate for each hypothetical criterion.

Finally, for each criterion, the corresponding hit rate and false alarm rates are plotted to form the C-ROC function. The C-ROC function corresponding to the data from Table B1 is shown in Figure B1.

Table B3
Cumulative Hit Rate and False Alarm Rate for Hypothetical Decision Criteria (k), for Both Signal and Noise Trials

	Criterion (k)						
	$k < 1$	$1 < k < 2$	$2 < k < 3$	$3 < k < 4$	$4 < k < 5$	$5 < k < 6$	$6 < k$
False alarms	100	82	54	27	20	8	0
Hits	100	95	83	73	48	23	0

APPENDIX B (Continued)

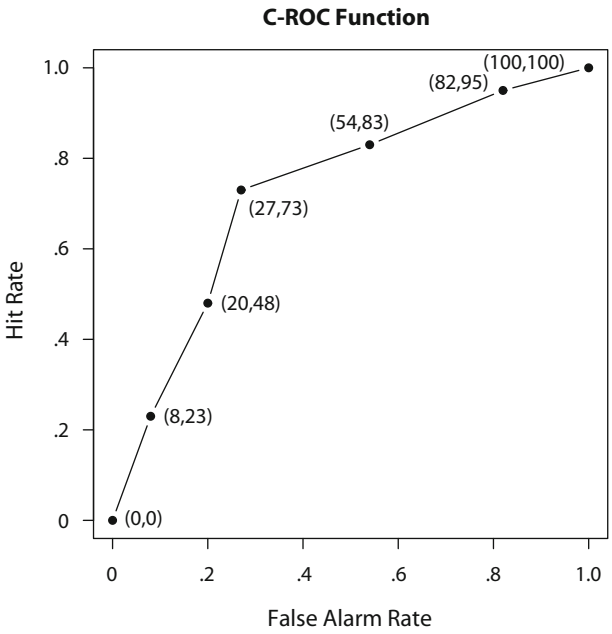

Figure B1. C-ROC function from hypothetical data in Table B1. Each point is identified with its corresponding numbers from Table B3, which are 100 times the plotted values.

Computing the DS-ROC Function

The DS-ROC function is computed by applying the same steps used to compute the C-ROC function, but on the columns of Table B1 rather than on the rows. First, the different confidence responses from Table B1 are aggregated into two classes ("high" vs. "low"), maintaining the distribution across distal stimulus intensities. These values are shown in Table B4.

Table B4
Number of Trials Across Stimulus Intensities for Each Response Class

	Response	
Stimulus Intensity	Low	High
100	16	7
200	32	12
300	25	8
400	11	21
500	10	26
600	6	26
Σ	100	100

Just as hit rate and false alarm rate were computed for the C-ROC function, one computes the distribution of stimulus classes for each transition between stimulus intensities. In effect, this procedure answers the question, "What are the probabilities of 'low' and 'high' ratings for different levels of stimulus intensity?" For example, if the hypothetical division between signal and noise were between 200 and 300, then this would have produced 52/100 signal trials for "low" responses and 81/100 signal trials for "high" responses. For each hypothetical point separating stimulus intensities, the number of "low" and "high" responses are shown in Table B5.

(*Continued on next page*)

APPENDIX B (Continued)

Table B5
Proportion of Trials With Intensity Greater Than *k*, for All Hypothetical Values of *k* and Both Response Classes

	Response	
Division (*k*)	Low	High
$k < 100$	100	100
$100 < k < 200$	84	93
$200 < k < 300$	52	81
$300 < k < 400$	27	73
$400 < k < 500$	16	52
$500 < k < 600$	6	26
$600 < k$	0	0

Finally, values in this table are converted to relative frequencies and plotted against one another to form the DS-ROC function. Figure B2 shows the DS-ROC for our example, with each point labeled by the corresponding numbers from Table B5.

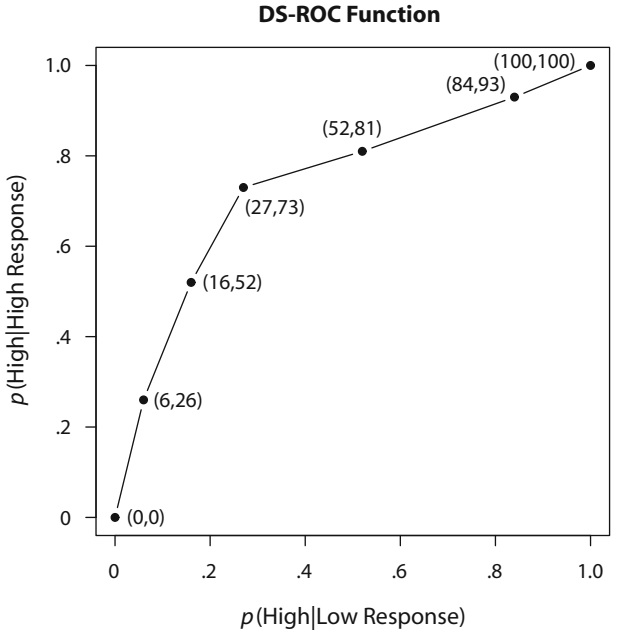

Figure B2. DS-ROC function from hypothetical data in Table B1. Each point is identified with its corresponding numbers from Table B5, which are 100 times the plotted values.

Application to Real Data

The contrived example we provided is convenient, because regardless of whether stimuli were divided into two classes on the basis of stimulus or response, there were exactly 100 of each. This is typically not the case for experimental data, but the analysis does not depend on the overall proportion of stimuli and responses being equal.

In many experiments, not all data shown in Table B1 are available. Typically, no independent measure of stimulus intensity is recorded; thus, when confidence ratings are collected, only the data in Table B2 are available. On the other hand, if one wants to compute the DS-ROC function, then no confidence ratings are required, and data collected in an experiment that did not use confidence ratings might look like those in Table B5. In our experiments, we used stimulus intensities that varied over a range of about 30 points, so our data matrix contained 30 rows and 8 columns. Because we have so many levels of stimulus intensity, we typically plot the DS-ROC as the line connecting the points, without identifying each individual point in the function.

(Manuscript received March 30, 2006;
revision accepted for publication September 25, 2007.)