Psychonomic Bulletin & Review
2008, 15 (5), 1022-1030
doi: 10.3758/PBR.15.5.1022

NOTES AND COMMENT

Decision criteria do not shift: Commentary on Mueller and Weidemann (2008)

J. D. BALAKRISHNAN
California Polytechnic State University, San Luis Obispo, California

AND

JUSTIN A. MACDONALD
New Mexico State University, Las Cruces, New Mexico

The effects of base rates and payoffs on the shapes of rating receiver operating characteristic curves are inconsistent with the basic assumptions of signal detection theory (SDT), in particular the notion of a shifting decision criterion. Mueller and Weidemann (2008) propose that these unexpected phenomena are not due to problems with the decision-criterion construct but are instead due to two compounded effects: instability of the decision criterion across trials, and even greater instability in the flanking criteria that determine which confidence rating will be reported. There are several problems with the authors' decision-noise hypothesis. First, even if their hypothesis about decision noise were taken for granted, the key feature of the ratings data that rejects the SDT model would remain a mystery. Second, the same violations of SDT that are exhibited in the ratings paradigm are also exhibited in the yes–no detection task when response time is substituted for confidence as a basis for analysis. Finally, the decision-noise hypothesis predicts that sensitivity will increase when one source of this variation—the response on a previous trial—is controlled for. This prediction was consistently violated in both the yes–no and ratings conditions of Mueller and Weidemann's experiment. In an Addendum, we respond to Weidemann and Mueller's (2008) reply to this Comment.

In many areas of perception and memory research, experimental phenomena that appear to have significant implications about perception or memory processes per se could actually be due to the effects of response biases. For the past 50 years or so, the most widely accepted method of distinguishing these two possibilities has been to apply a signal detection analysis to the data (e.g., Green & Swets, 1966). Signal detection theory (SDT) is, at its core, the formal expression of an intuitively compelling idea—that is, that decisions about stimuli are based on decision criteria. A biased decision criterion demands relatively stronger sensory or memory evidence before an unpreferred response will be prescribed. The signal detection model is thought to be well supported by scores of classical studies, followed up by many years of apparently successful applications.

Recently, we reported two empirical results that appear to be as robust as any of the classical results in the SDT literature and yet are inconsistent with the classical SDT framework (Balakrishnan, 1998a, 1998b, 1999; see also Van Zandt, 2000). The first problem is the fact that the receiver operating characteristic (ROC) curves obtained from *yes–no* detection experiments with confidence-rating responses change shape under different biasing conditions. The second problem is that the likelihood-ratio function (a measure that is closely related to the ROC curve, as we explain in this article) is always very close to 1 at the point of the function corresponding to the lowest confidence responses. Both properties are anathema from the SDT point of view. The first result contradicts the assumption that the sensory or memory effects of the stimuli do not depend on the amount of response bias in the decision process; the second result contradicts the assumption that a change in response bias is a change in how sensory or memory states are mapped to responses—that is, a shift in the decision criterion.

Three attempts to account for these empirical phenomena, without giving up on the central concepts of SDT, have been published: Treisman (2002), Kornbrot (2006), and Mueller and Weidemann (2008). Treisman's and Kornbrot's arguments were inadequate, in our view, for various reasons, the most obvious being that they did not show that any kind of signal detection model could actually fit the data we reported. Mueller and Weidemann were the first to propose an extension of the classical detection model that appeared to account for the observed results, and did so without dropping the crucial idea of the decision criterion. They suggested, first, that there is a substantial amount of variation from trial to trial in the location of the decision criterion and, second, that there is even more variation in the locations of the additional criteria that determine the reported degree of confidence in the response. The difference in the degree of variation between the two types of criteria causes the shape of the rating ROC curve to be deformed in different ways, depending on the average placement of the decision criterion. Mueller and Weidemann also pointed out, however, that if they are correct about the source of the violations of the classical signal detection model, the SDT statistics (e.g., d' and β) would not adequately distinguish sensory or memory effects from response biases, as most psychologists have assumed that they do. The authors therefore suggested that more sophisticated methods of analysis are long overdue.

In this article, we show that allowing for the possibility of two different kinds of decision noise, although it may be plausible from an SDT perspective, is not enough to explain the incorrect predictions of classical SDT that we reported, for two main reasons. First, in order to fit the

J. D. Balakrishnan, jbalakri@calpoly.edu

ratings data, the ratio of decision-criterion noise to rating-criteria noise must always fall within the small range of values needed to cause the likelihood-ratio function to approach the value 1 as confidence decreases, instead of any other value. Because the specific value of this ratio has no special meaning in Mueller and Weidemann's (2008) theory, their model would fit the data without offering a meaningful explanation of them. Second, if the incorrect predictions of SDT are merely due to the manner in which participants make confidence ratings, as Mueller and Weidemann suggested, there is no reason to expect the same problematic results to be observed when, instead of the rating responses, participants make only the *yes*-or-*no* detection response, and response time (RT) is substituted for the degree of subjective confidence in order to compute likelihood-ratio functions. When participants make only a *yes*-or-*no* response, there is only one decision parameter—the decision criterion—in the signal detection model, and trial-by-trial variability in this criterion alone, however large it might be, is not sufficient to account for the behavior of the rating ROC and likelihood-ratio functions. Mueller and Weidemann were careful to include in their experiment a standard *yes–no* detection condition in addition to their ratings condition, and they also recorded RTs. This gave us the opportunity to perform all of the same tests on RT data that we previously applied to ratings data. The results confirmed that the violations of SDT are not merely due to decision noise or to the manner in which participants make confidence ratings.

Most psychologists are familiar with the basic principles of SDT, including the concept of the decision criterion, the ROC curve, and, at least to some extent, the use of confidence ratings to estimate an ROC curve. The theoretical significance of the likelihood-ratio function in SDT is not as well known, and empirical estimates of these functions are almost never reported. However, more than the ROC curve, or even the decision criterion itself, it is the likelihood-ratio function that determines whether a decision maker is biased in the manner envisaged by SDT. Before considering Mueller and Weidemann's (2008) hypothesis about decision noise, therefore, we first explain how the likelihood-ratio function and the notion of response bias are connected. We then show how ratings and RT data can be used to determine whether the likelihood-ratio function behaves in a manner consistent with SDT.

Detection Theory: Background and Terminology

In two-choice classification tasks (e.g., *yes–no* detection, discrimination, recognition memory), the relative frequency of the *A* response on *a* trials (the correct-rejection rate) will usually be different from the relative frequency of the *B* response on *b* trials (the hit rate), even if the two stimulus types are presented equally often (the base rates are equal). Most psychologists take it for granted that this difference is generally due to a response bias of some kind. Because the direction and size of this difference can change in relatively unpredictable ways in different circumstances, response biases could easily mislead an investigator if their effects on behavior are not somehow accounted for.

The nature of a response bias in the SDT framework is illustrated in Figure 1. Values on the *x*-axis of the figure represent the sensory state (or the familiarity of a recognition memory test probe) that results from encoding the stimulus. The distribution on the left, f_a, represents the relative frequencies of the sensory states on *a* trials, whereas the distribution on the right, f_b, represents the relative frequencies of the sensory states on *b* trials.[1]

The exact shape of the distributions (normal or otherwise) is not a central concept in SDT. What is important is that they overlap to some degree. That is, their heights are unequal and greater than 0 for at least some values on the *x*-axis (i.e., for at least some sensory states). Because of this overlap, the sensory states provide partial, but not perfect, information about the stimulus type on each trial, making classification errors unavoidable. The observer's solution to this *statistical decision problem* is to choose a decision criterion, X_C, to divide the sensory states into *A* and *B* responses, with each sensory state to the left of the criterion being mapped to an *A* response and each state to the right being mapped to a *B* response. The mapping of each possible sensory state to one and only one classification response is the observer's *decision rule*.

The predicted correct-rejection rate in SDT is equal to the area under the f_a distribution to the left of the decision criterion. The area to the right of the criterion under the same distribution is the false alarm rate. The hit and miss rates are the areas under the f_b distribution to the right and left of the decision criterion, respectively. In the Figure 1 example, the distributions are symmetric and identical except for their means, −1 and 1, and they intersect only once, when the sensory state value is equal to 0. The decision criterion, X_C, is indicated by the long vertical bar, which falls to the right of this intersection point. The predicted hit rate in this example is therefore less than the predicted correct-rejection rate, due to the placement of the decision criterion.

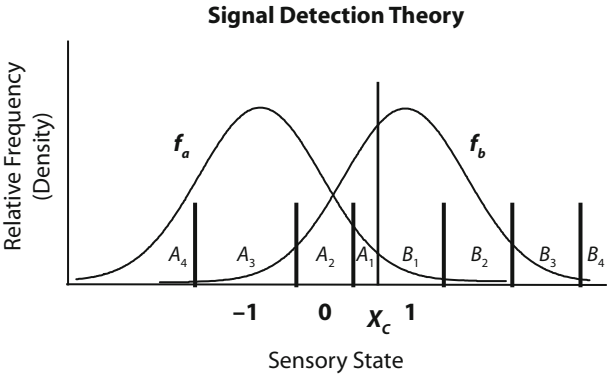

Figure 1. The classical signal detection model of ratings and discrimination behavior. The sensory effect of the stimulus is represented by a single real number (the sensory state), which is mapped to one of the two possible classification responses by choosing a decision criterion, indicated by the long vertical bar. The additional, rating criteria (short vertical bars) further partition the sensory states into distinct confidence-rating response regions, with confidence increasing as the sensory state approaches one of the extremes of the sensory state dimension.

In exactly what sense is this a "biased" decision rule? Although it is true that the decision criterion is shifted to the right of the midpoint, 0, between the means of the two distributions, the notion of bias in SDT is defined (for good reason) in a more technical manner. Specifically, in the region of sensory states between the value 0 and X_C, the height of the f_a distribution is less than the height of the f_b distribution. Yet, these states will be mapped by the observer to the *A* response. In other words, there is a *biased-response region*, in which the sensory information favors the *B* response, but the observer responds *A*. In SDT, the decision rule is biased if and only if there are one or more response regions of this kind (i.e., one distribution is higher than the other, but the classification response corresponds to the shorter distribution). A biased decision rule is not necessarily a poor decision rule: If the base rates are unequal, it would usually be necessary to bias the decision rule in order to maximize accuracy.

The Rating ROC Curve

The basic rationale behind the decision criterion parameter in SDT is that the observer recognizes that the sensory state dimension is a *bipolar* scale. That is, the stimulus is increasingly likely to be of type *a* as the sensory state decreases, and increasingly likely to be of type *b* as the sensory state increases, and the observer is aware of this statistical fact. Somewhere on the sensory dimension is a point where the probabilities of the two stimuli are equal (the *A* and *B* responses each have a 50% chance of being correct). When the base rates are equal, this would be the 0 point in Figure 1, where the two sensory distributions intersect, and the sensory information is therefore perfectly ambiguous. An observer whose objective is to be correct as often as possible will attempt to place the decision criterion at the point where the sensory information, combined with the base rates, makes the two possible stimuli equally likely.

When observers are asked to make confidence-rating responses in addition to their classification response, SDT assumes that additional rating criteria are placed on the sensory state dimension, some above and some below the decision criterion, in order to convert (map) each sensory state to one and only one rating response. The short vertical bars in Figure 1 illustrate the SDT model for the case in which there are four levels of confidence on the rating scale. Like the decision criterion, each rating criterion in Figure 1 has associated with it a pair of hit and false alarm rates, and these values can be estimated from the (cumulative) relative frequencies of the rating responses, without requiring any specific assumptions about the shapes of the sensory state distributions (see, e.g., Green & Swets, 1966). Plotting the hit rates against their corresponding false alarm rates yields the so-called rating ROC curve.

The shape of the underlying ROC curve—that is, the curve generated by calculating the hit and false alarm rates for *every* sensory state—will depend on the shapes and the degree of overlap of the sensory state distributions. As Mueller and Weidemann (2008) noted, the placement of the criteria will merely determine which of these points on the complete underlying ROC curve are plotted in the rating ROC curve. Therefore, shifting some or all of the criteria by some arbitrary amount in Figure 1 will change the hit and false alarm rates, but the new ROC points generated will fall on the same underlying function. If the two sets of points obtained from two different experimental conditions do not fall on a single underlying curve, changes in criteria positions alone would not be enough to fit the data from the two conditions: The sensory state distributions must also be allowed to change in some way.

Empirical rating ROC curves obtained under conditions that are presumed to differ only in the degree or direction of the response bias—that is, under conditions that differ only in the base rates of the stimuli or the payoff scheme—do not fall along a single underlying curve (Balakrishnan, 1998a, 1999; Van Zandt, 2000). As the difference between the hit and correct-rejection rate increases, the peak of the ROC function with respect to the negative diagonal shifts, and the degree of skewness of the curve increases. The effect is strong enough to make the assumption that the sensory state distributions are invariant clearly unreasonable, even as a first approximation. The SDT model can therefore be rejected. However, the assumption in SDT that the sensory distributions are independent of the amount of response bias is logically distinct from the model's assumption about the nature of the response bias itself. In principle, the shapes of the distributions could depend on the amount of bias, whereas the difference between the hit and correction rates could still be due to the presence of a biased-response region in the decision rule (i.e., to the placement of the decision criterion). To rule this possibility out, a different kind of analysis, involving the likelihood-ratio function, is needed.

The Empirical Likelihood-Ratio Function

If the sensory state on each trial of the classification experiment was observable, the presence of the biased-response region in the Figure 1 example would be easy to detect. Suppose that a very large experiment is performed and the stimulus (*S*), the classification response (*R*), and the sensory state (*I*) are recorded on each trial. Let *v* be any value that falls in the biased-response region in Figure 1. In this hypothetical data set, the classification response would always be $R = A$ whenever the sensory state *v* is recorded on a given trial, and the relative frequency of this observed state on *b* trials,

$$g_b(v) = \frac{\text{freq}(I = v, S = b)}{\text{freq}(S = b)},$$

will be found to be greater than the relative frequency of this state on *a* trials,

$$g_a(v) = \frac{\text{freq}(I = v, S = a)}{\text{freq}(S = a)}.$$

In fact, the ratio of these two empirical values—the empirical likelihood ratio, $g_b(v)/g_a(v)$—would be equal to the ratio of the two sensory state distributions, f_b to f_a, at the value *v* in Figure 1—that is, $f_b(v)/f_a(v)$. Thus, in this hypothetical experiment, the observable likelihood ratio associated with a certain kind of *A* response—that is, an *A* response that is accompanied by the sensory state *v*—would be found to be greater than 1, establishing the presence of

a biased-response region in the mapping of sensory states to classification responses.

Although the sensory states themselves are not observable, the presence of a biased-response region would still be detectable in some cases, depending on the placement of the rating and decision criteria. In fact, the Figure 1 model is an example of such a case. Due to the position of the rating criterion that falls immediately to the left of the decision criterion, the observer will only respond *A* at the lowest confidence level (henceforth, rating response A_1) when the sensory state falls somewhere inside the biased-response region. Therefore, whenever the observer makes this particular rating response, the likelihood ratio associated with the incident sensory state must have been a value greater than 1. In such a case, the observable likelihood ratio associated with this rating response,

$$L(A_1) = \frac{\dfrac{\text{freq}(R = A_1, S = b)}{\text{freq}(S = b)}}{\dfrac{\text{freq}(R = A_1, S = a)}{\text{freq}(S = a)}},$$

will be found to be greater than 1 (if estimated from a sufficiently large data set).

Stated in more general terms, the observable likelihood ratio corresponding to any given rating response will always be a weighted average of unobservable likelihood ratios—that is, the likelihood ratios that accompany the sensory states that are mapped to the given rating response. The likelihood ratios that occur more frequently when a given rating response is reported will be weighted more strongly (see Balakrishnan, 2006). In fact, the weight assigned to each likelihood ratio is simply the relative frequency with which this (unobservable) ratio occurs when the (observable) rating response is selected by the observer.

The weighted average of a set of numbers cannot be less than the lowest value in the set of numbers to be averaged, or greater than the greatest value in the set. It is impossible, therefore, for the (observable) likelihood ratio associated with a given rating response to be greater than 1 if each of the unobservable likelihood ratios that are mapped to this response is less than 1. Therefore, showing that the empirical likelihood ratio is greater than 1 for one or more *A* rating responses, or is less than 1 for one or more *B* rating responses, establishes the presence of a bias in the decision rule (i.e., a response bias of the SDT kind) in the strongest possible terms.

Of course, it is possible for the rating criterion that falls immediately to the left of the decision criterion in Figure 1 to be shifted far enough to the left so that not all of the likelihood ratios in the set of sensory states mapped to the A_1 rating response are greater than 1. If the spacing of the criteria is large enough, the relative frequency of these small (less than 1) likelihood-ratio values when rating response A_1 is reported will be large enough to cause the weighted average to be less than 1 [i.e., $L(A_1)$ will be less than 1]. This would happen, for example, if the rating criterion immediately to the left of the decision criterion in Figure 1 were shifted to, say, a point somewhere to the left of the mean of the f_a distribution.

However, the spacing between the criteria in the Figure 1 model not only determines which sensory states (and hence which likelihood ratios) are mapped to which rating responses, it also determines how often a given rating response will be selected. If the observed relative frequency of the A_1 response is very small, the spacing between the two adjacent criteria that determine when this rating response will be selected cannot be very large. As the spacing in the Figure 1 example is reduced, the likelihood ratio associated with response A_1 must approach the likelihood ratio at the decision criterion, which is greater than 1.

In the visual discrimination experiments that we reported, the relative frequency of the lowest confidence *A* response was very small (e.g., less than 2%), but the likelihood ratio associated with this response was close to 1 (e.g., .98), even when the base rates were very different (9 to 1 in favor of the *A* response), causing the correct-rejection rate to be more than 30 percentage points greater than the hit rate (Balakrishnan, 1998a).[2]

The Decision-Noise Model

To explain the unexpected behavior of the rating ROC curves and the rating likelihood-ratio function, we pointed out that both results are predicted, without requiring any ad hoc assumptions, by sequential-sampling models of classification (Balakrishnan, 1999; see also Van Zandt, 2000). In their alternative proposal, Mueller and Weidemann's (2008) focus was on the decision process in SDT and its effect on the shape of the rating ROC curves. In SDT, the decision process is the mapping of sensory states to observable responses—that is, the placement of criteria. In their decision-noise model, Mueller and Weidemann assumed, first, that the rating and decision criteria shift from trial to trial and, second, that the rating criteria are even more unstable than the decision criterion. They showed that under these two assumptions the rating ROC curves obtained from two conditions that differ only with regard to the placement of criteria (the sensory state distributions are invariant) will be different in shape, in a manner consistent with the empirical rating ROC curves. Although they did not report likelihood-ratio functions, the authors indicated that their model was also able to adequately reproduce the frequencies of the lowest confidence rating responses, even when these were very small.

The hypothesis that decision criteria vary and that rating criteria are even less stable than decision criteria seems plausible enough and does appear to offer a relatively simple reason why base rate and payoff manipulations should be expected to change the shapes of the rating ROC curve. However, this hypothesis does not explain the behavior of the rating likelihood-ratio function. When the positions of the rating and decision criteria vary from trial to trial, the observable likelihood ratio associated with a given rating response will still be a weighted average of unobservable likelihood ratios, and the weights involved in the average will still be the relative frequencies with which the unobservable ratios occur when the given rating response is selected. The only difference between fixed and variable criteria is that in the latter case the set of likelihood ratios that sometimes occur when a given rating response is selected (i.e., the

likelihood-ratio values that have nonzero relative frequencies) is larger, due to the variation in the criteria from trial to trial. A sensory state that on one trial might be far away from the region of states mapped to rating response A_1 could fall within this region on another trial, when this region is defined by a different pair of upper and lower boundaries. In this respect, the effect of decision noise on the observable likelihood-ratio function is similar to the effect of increasing the spacing between the criteria in the SDT model.

To illustrate, suppose that the rating and decision criteria in Figure 1 are merely the average values of the criteria, instead of fixed constants that describe every trial. If the variability in the rating criterion immediately to the left of the decision criterion is large enough, this criterion will sometimes fall to the left of the point of intersection (0) between the two distributions. On some of these trials, the sensory state will also be less than 0, but will still fall inside the A_1 response region. Thus, the unobservable likelihood ratio when the observer responds A_1 will be less than 1. If the average values of the rating and decision criterion are sufficiently close together, their spacing on a given trial would often be small, and the relative frequency of the lowest confidence rating response could therefore turn out to be small, despite the fact that the range of sensory states that are at least sometimes mapped to this response can be quite large.

Considered in these general terms, it seems possible that by adjusting the average values of the criteria and the different amounts of rating- and decision-criterion noise, the decision-noise model (or some minor variation of it) could produce likelihood-ratio functions that approach the value 1 as confidence decreases, even when there is a substantial difference between the predicted hit and correct-rejection rates. However, it is important to recognize that the value 1 would have no special significance in this model: The model does not predict that the likelihood-ratio function *should* always approach the value 1, it merely accommodates this empirical fact post hoc by adjusting the ratio of rating-criteria noise to decision-criterion noise so that, from the wide range of possible results, the value 1 becomes a coincidental constant. Thus, even if the authors' decision-noise model is assumed to be accurate in every detail, the behavior of the empirical likelihood-ratio function would still be a mystery.

The RT Likelihood-Ratio Function

Mueller and Weidemann's (2008) main purpose in developing their decision-noise model was to support their thesis that the problems that we documented with SDT could "stem from the confidence rating procedure itself" (p. 467). In this section, we show how a likelihood-ratio function can be computed from RT data instead of ratings data, making it possible to estimate these functions when rating responses are not even solicited from the participants.

In the SDT model for the ratings experiment, the presentation of the stimulus gives rise to a sensory state (v), a corresponding likelihood ratio (l_v), and a classification response (A or B) that has a confidence level (k) assigned to it. Although there is no mention of it in Figure 1, the RT could also be recorded on each trial. If the participants are not asked to make rating responses, as in the classical *yes–no* detection task, there is still a decision criterion, but there are no rating criteria in the SDT model; the short vertical bars in Figure 1 would be superfluous. However, even when there are no rating responses, the classification response on each trial will still have an RT value assigned to it. In the same way that there are different types of A responses in the ratings paradigm, depending on what degree of confidence is reported, there are different types of A responses in the *yes–no* experiment, depending on the time taken to make this classification response. The RT likelihood ratio is the relative frequency with which the classification response will be X (where $X = A$ or B) and the RT will be t on b trials, divided by the relative frequency with which the classification response will be X and the RT will be t on a trials,

$$L\left(X_{\mathrm{RT}=t}\right) = \frac{\dfrac{\mathrm{freq}\left(R = X_{\mathrm{RT}=t}, S = b\right)}{\mathrm{freq}(S = b)}}{\dfrac{\mathrm{freq}\left(R = X_{\mathrm{RT}=t}, S = a\right)}{\mathrm{freq}(S = a)}},$$

where $X = A$ defines the left side of the function, and $X = B$ defines the right side of the function.

At least as it is generally understood, SDT does not make specific predictions about the relationship between sensory states and RT. However, because on each trial there is an RT and an incident likelihood ratio, the same theorem that shows that observable rating likelihood ratios are weighted averages of unobservable likelihood ratios also applies to the RT likelihood-ratio function. For example, if the (unobservable) likelihood ratio, l_v, is always equal to 2 whenever the time to respond A is 1.3 sec, the experimenter will find that the observable statistic, $L(A_{\mathrm{RT}=1.3})$, will be equal to 2 (if computed from a sufficiently large sample). If l_v is 2 on 25% of the trials when the response is A and the RT is 1.3 sec, and l_v is 4 on the remaining 75% of these trials, the RT likelihood ratio, $L(A_{\mathrm{RT}=1.3})$, will be found to equal $\frac{1}{4} * 2 + \frac{3}{4} * 4 = 3.5$.

The main reason for estimating the RT likelihood-ratio function is to see whether it copies the behavior of the rating likelihood-ratio function. If the Figure 1 model is accurate, but there is no dependence whatsoever between the observable RT and the unobservable likelihood ratio, then breaking the classification responses down into different RT categories will be useless: The estimated RT likelihood-ratio function will be a (noisy) step function, having one underlying value when the classification response is A and another value when the classification response is B. However, it is well known that RT and accuracy are usually correlated; fast responses are typically more likely to be correct than slow responses (see, e.g., Katz, 1970; Link, 1992). To explain this relationship, there must be at least some correlation between the unobservable likelihood ratio and RT.

Statistical issues. Unlike a confidence-rating scale, which would typically subdivide the classification responses into no more than 10 different confidence levels, the RTs recorded by the experimenter will assume many different values, especially if they are recorded with high (e.g., millisecond) accuracy. The true RT likelihood-ratio

function will therefore have many distinct points, and each of these points would need to be estimated from an extremely small sample. However, the recorded RT value on a given trial of any real experiment will always represent an interval of actual RT values—that is, an RT bin, which contains the true value that actually occurred. In the same way that large confidence-rating scales can be reduced in size by merging adjacent levels on the rating scale, the number of different RT values can be reduced by combining adjacent RT scores. In fact, there is no reason why the sizes of the RT bins must be constant; they can be chosen so that there will be an equal number of samples in each bin, or so that the smallest sample size is still large enough to provide a reasonably accurate estimate of the likelihood ratio.

Results

There is a close relationship between the shape of an ROC curve and the shape of the likelihood-ratio function (the likelihood-ratio function is the first derivative, or rate of change, of the ROC curve; see, e.g., Green & Swets, 1966). However, apart from the question of whether two data sets can be fit with a single pair of sensory distributions with different rating criteria (as discussed above), the shape of the ROC curve is a relatively weak basis for testing hypotheses about decision processes (cf. Balakrishnan, MacDonald, & Kohen, 2003). Our focus is therefore on the likelihood-ratio functions.

To estimate RT likelihood-ratio functions for the *yes–no* condition of Mueller and Weidemann's (2008) experiment, we defined the RT bins so that each bin contained 300 samples. To achieve this, we sorted the data by RT from largest to smallest and then let each successive block of 300 scores define a different RT bin. For example, in the high *a* base rate condition of Mueller and Weidemann's experiment, the slowest RT was 28 sec and the 300th slowest RT was 2.2 sec. The slowest RT bin was therefore defined by the interval from 2.2 to 28 sec. Because there were 12,000 trials in each base rate condition, there were 12,000/300 = 40 different RT bins, and therefore 80 different points (40 for each classification response) on the RT likelihood-ratio function.

The results for the two different base rate conditions are shown in Figure 2. The dashed vertical bar in the two panels of Figure 2 indicates the point at which the classification response switches from *A* to *B* (the two sides of the function). The dashed horizontal bar indicates the crucial point at which the likelihood-ratio function reaches the value 1. In both conditions, the functions are less than 1 on the *A* response side and greater than 1 on the *B* response side. More important, the function comes very close to the value 1 as the RT decreases (the middle of the graph). Technically, it approaches 1 again as RT increases in one case (the right edge in the lower panel), but since this is due to a single point, the result could be simply the result of estimation error.

In the Figure 1 model, the unobservable likelihood ratios to the right of the decision criterion are all greater than 1, with the lower bound being the likelihood ratio at the decision criterion. According to this model, therefore, the RT likelihood-ratio function cannot reach the value 1 on the *B* response side, regardless of how the unobservable likelihood ratios and RT are related. In fact, it cannot fall below the likelihood ratio at the decision criterion. In both of the base rate conditions, the functions reach the value 1 on both sides. To illustrate the size of the discrepancy, in the high *b* base rate condition (lower panel), the hit rate was almost 30 percentage points larger than the correct-rejection rate, and the SDT bias parameter β—that is, the supposed value of the (unobservable) likelihood-ratio function at the point of the decision criterion—was 0.72. Therefore, the value 0.72 should have been the largest value reached by the RT likelihood-ratio function on the *A* response side in the lower panel of Figure 2, and the smallest value reached on the *B* response side. In crude terms, the results show that on at least some trials in the high *b* base rate condition, the participants respond *A* even when their sensory evidence favors this response to a trivially small degree (the likelihood ratio is close to or equal to 1). In more technical terms, their decision rule appears to be unbiased in the SDT sense, despite the difference in the hit and correct-rejection rates.[3]

Implications of Reverse Speed–Accuracy Trade-Offs

The fact that participants do not follow the prescriptions of SDT when they have little or no sensory information to go by is not the only inconsistency revealed by the estimated likelihood-ratio functions in Figure 2. Equally problematic for SDT is the fact that the likelihood ratio (and hence, accuracy) approaches the value 1 (chance performance) as RT increases from moderately fast to very slow. In other words, ignoring the very fast responses, there is a negative correlation between RT and accuracy, as is typical of discrimination experiments. This reverse speed–accuracy trade-off is modulated by the base rates: The likelihood ratio tends to be further away from the value 1 on the less preferred response side of the RT likelihood-ratio function. For example, in the lower panel of Figure 2 (high *b* condition), the highest value reached by the likelihood-ratio function on the *B* response side was roughly 2.5, or 2.5 to 1, whereas the lowest value reached on the *A* response side was roughly 0.2, or 1 to 5.

The reverse speed–accuracy trade-off is contrary to the basic principles of SDT and, in particular, its assumption that decision processes are merely rules for assigning responses to sensory states. If anything, there should be a positive correlation between RT and accuracy: Taking more time to deliberate should allow the participant to collect more, not less, sensory information before the decision process intervenes. This particular misprediction of SDT is one of the strongest endorsements of sequential-sampling theories and their alternative interpretation of the nature of a response bias. In these models, the negative correlation between RT and accuracy and the effects of base rates on this relationship are both the expected consequences of a single decision mechanism. In rough terms (i.e., see Balakrishnan, 1999, and Van Zandt, 2000, for more detailed discussions), sequential-sampling models assume, first, that the participants are concerned about response speed as well as accuracy and, second, that they are able to interrupt the sensory encoding process when

it seems prudent to do so. If, after a short deliberation, the evidence collected so far is already strongly in favor of, say, the *A* response, then there is little to be gained (in accuracy) by continuing to deliberate, and the participant therefore responds immediately, forfeiting the opportunity to collect more evidence.

By stopping the encoding process quickly only when the evidence so far accumulated is relatively conclusive, the decision maker forces the fast responses to be relatively more accurate. For the same reason, the slow responses end up absorbing all of the trials in which the evidence never became very conclusive in either direction, resulting in less accuracy on these trials and, hence, a decrease in accuracy as RT increases. The effects of the base rates on this kind of decision-making strategy are more difficult to illustrate. However, the most important step in the analysis is simple enough. When the base rate of, say, the *a* stimulus is greater than the base rate of the *b* stimulus, the amount of sensory evidence in favor of the *A* response that would be needed to reach the point at which further deliberation would be too costly to be justified is less than the amount in favor of the *B* response that would be needed to reach this same point. The asymmetry in the two sides of the RT likelihood-ratio function is one of several consequences of this asymmetry in the standards of evidence applied by the participants when they decide when it is appropriate to stop deliberating and respond.

Effects of Criterial Noise on the Sensitivity Level

To support their hypothesis that decision noise contributes substantially to the overt performance of participants in discrimination tasks (enough to explain the violations of SDT that we reported), Mueller and Weidemann (2008) pointed out that there are sequential effects in a typical classification experiment. One well-documented effect of this kind is the tendency for the classification response given on a previous trial to be repeated on the subsequent trial. In SDT terms, the hit and false alarm rates increase immediately following a *B* response and decrease immediately following an *A* response (see, e.g., Parducci & Sandusky, 1965; Treisman & Williams, 1984). Following an *A* response, the decision criterion presumably tends to fall to the right of its overall average position, and following a *B* response, it tends to fall to the left of its overall average position. In other words, the decision criterion appears to shift from

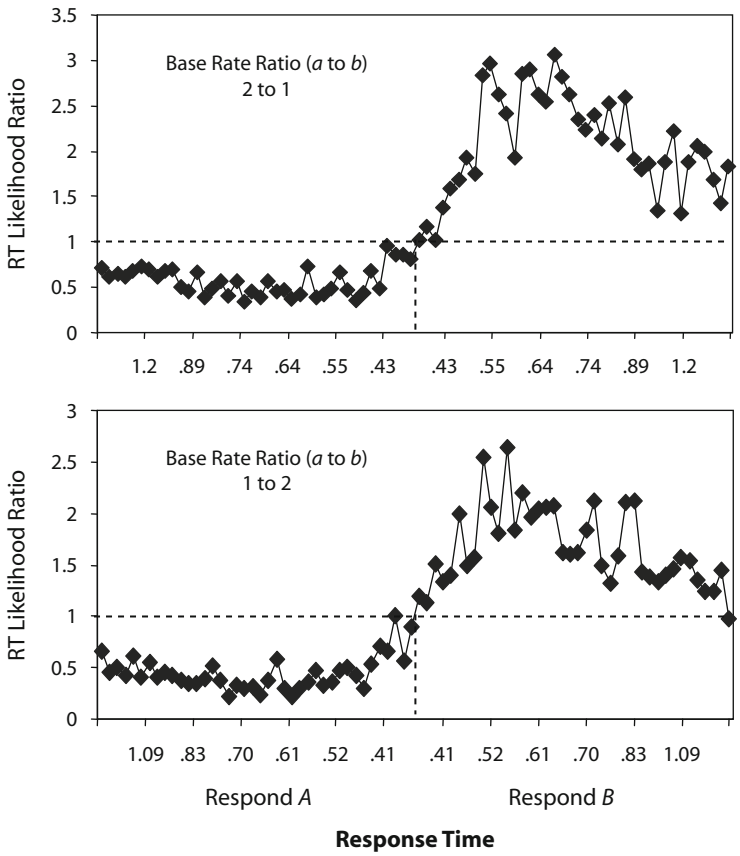

Figure 2. The response time (RT) likelihood-ratio functions corresponding to the RT-ROC curves shown in Figure 1 (the likelihood-ratio values are the slopes of the ROC curve). The dashed vertical bar defines the two sides of the function—that is, the point separating the *A* and *B* responses. The dashed horizontal bar indicates the crucial value 1, which determines whether the decision rule is biased or unbiased. Contrary to the signal detection theory prediction, the functions approach the value 1 on both sides as RT decreases.

Table 1
Effects of Controlling for the Prior Response on the Estimated Sensitivity, *d′*, in the Six Conditions (2 Response Procedures × 3 Base Rates) of Mueller and Weidemann's (2008) Experiment

	Base Rate Ratio (*a* to *b*)		
Prior Response	1 to 1	1 to 2	2 to 1
Ratings			
A	.780	.812	.836
A or *B*	.805	.824	.793
B	.842	.854	.755
Yes–No			
A	.732	.888	.783
A or *B*	.767	.833	.793
B	.794	.816	.838

trial to trial, as in Mueller and Weidemann's decision-noise model.

As Mueller and Weidemann (2008) pointed out, adding noise to the decision-criterion parameter in the SDT model has the same effect on performance as keeping the decision criterion constant and adding additional noise to the sensory state distributions (i.e., reducing the sensitivity level). Controlling for the response on a previous trial, therefore, should reduce the effect of criterial noise and, accordingly, increase the estimated sensitivity level.

Table 1 compares the overall *d′* value (i.e., the prior response is *A* or *B*) to the two sequence-dependent *d′* values for the six different conditions of Mueller and Weidemann's (2008) experiment. In each case, the overall *d′* value falls between, instead of below, the two sequence-dependent scores. Thus, the effect of the previous response on the hit and correct-rejection rates unequivocally rules out the possibility that the decision criterion is constant across trials, consistent with Mueller and Weidemann's hypothesis; but simply allowing for trial-by-trial changes in the placement of the decision criterion is not enough to explain the effect of the previous response on performance. Since the same violation appears in both the ratings and *yes–no* detection conditions, it cannot be attributed to different degrees of noise in rating and decision criteria, or to any other artifactual consequence of soliciting confidence ratings in addition to a *yes*-or-*no* response.

Discussion

The idea that human observers will adopt different decision criteria under different circumstances, sometimes capriciously, sometimes in a strategic manner, seems to offer a simple and intuitively compelling way to explain a very simple fact—that is, that the two kinds of correct responses in a two-choice classification task, hits and correct rejections, usually do not occur with equal frequency, even if the two types of stimuli are presented equally often. SDT is a rigorous means of quantifying this basic intuition. However, the concept of the decision criterion in SDT is considerably less plausible than most psychologists probably realize. The signal detection model is, in effect, the solution to a specific kind of statistical decision problem, in which the decision maker is given some data and is then asked to choose a response. The theory implicitly assumes, therefore, that the decision maker has no control over the amount or quality of the information collected prior to selecting a response. If the decision maker prefers one response to another for some reason, the only manner in which this preference can be expressed is to adjust the decision criterion—that is, the mapping of information states to responses (the decision rule). Mueller and Weidemann's (2008) modified detection model is more flexible and more plausible in some important respects, but it also assumes that the participants' decision strategies are constrained in this way.

In the real world, there is always an incentive to respond promptly, as well as accurately. Otherwise, there would never be any reason to make risky decisions of any kind. This is true of expert decision problems, such as those faced by medical practitioners (more tests can be ordered before an intervention is recommended), as well as the typical kinds of sensory and memory discrimination problems defined by laboratory experiments. If the time to reach a final decision is even a small factor in participants' minds, the decision rule (and, hence, the decision criterion) is no longer an appropriate basis for describing the nature of the participant's decision-making strategy, or for distinguishing sensitivity effects from response biases.

The solution to this more realistic statistical decision problem is the sequential-sampling model (e.g., Link & Heath, 1975; Luce, 1986; Ratcliff & Smith, 2004). Although there are many specific examples of these models, and they can be difficult to distinguish empirically, they share the key assumption that observers are motivated to respond quickly, as well as accurately. Instead of changing the mapping of sensory states to responses, response biases in these models become properties of the *stopping rule*, which determines when the deliberation process will be terminated. The time spent deliberating determines, in turn, the amount of sensory information that is available to the decision maker at the point at which the response is finally selected. Described in crude terms (see Balakrishnan, 1999, and Van Zandt, 2000, for a detailed discussion), a response bias in this alternative framework is a difference in the strength of the sensory evidence needed to cause the decision maker to terminate the encoding process and respond. A decision maker who prefers the *A* response will stop deliberating and respond quickly if the sensory evidence collected early on favors this particular response. If the initial evidence favors the unpreferred response, the decision maker is likely to wait longer before responding, and during this extended period, the balance of evidence may switch to the preferred response. The preferred response is therefore more frequent and faster, consistent with the discrimination literature.

Like SDT, the sequential-sampling models cannot be used to quantify bias and sensitivity without introducing some specific, relatively technical assumptions about sensory noise and the decision process, and some of these assumptions are almost surely incorrect. However, these models are not only able to fit both RT and response frequency data with suitably chosen parameters, they are also constrained to predict (and, hence, they explain) features of RT data that, at least as much as response biases, one would be foolish to ignore, including the correlation between the means and variances of the RT distributions and

the relationship between response accuracy and response speed (i.e., speed–accuracy trade-offs). In this respect, sequential-sampling models seem to be a much more suitable starting point for the analysis of classification data, even when RT itself is not a measure of particular interest in a given study.

ADDENDUM

Reply to Weidemann and Mueller (2008)

In their reply to this commentary, Weidemann and Mueller (2008) have suggested, first, that confidence ratings and RT are both "noisy indices" of perceptual evidence. This hypothesis is self-evident, but it does not resolve any of the issues. To actually fit data, they introduce an assumption in which the only reason RTs are correlated at all with accuracy is that incorrect responses have larger RT variances than correct responses do. They offer no theoretical explanation for this RT variance assumption. Furthermore, their idea about two types of decision noise, and even the notion of a biased decision criterion, played absolutely no role in this "illustrative analysis." Even this "quasi"-model yielded only a partial fit to the data, reproducing the approach to the value 1 only on the "favored response" side of the RT likelihood ratio function. Thus, the authors never explained why participants who are presumably biased toward the "A" response always respond "B" when their sensory information slightly favors the "B" response. Reanalyzing data from Van Zandt (2000), the authors then pointed out that confidence-rating likelihood-ratio functions do not always approach the value 1. However, Van Zandt did not control the frequencies of the lowest confidence-rating responses. For the reasons detailed in this article (see p. 1025), this makes it impossible to estimate the value approached by the underlying function. If there is anything approaching a reasonably plausible SDT account of discrimination behavior, it has yet to be discovered.

AUTHOR NOTE

We thank Ehtibar Dzhafarov for helpful discussions and comments on a previous version of this article. We also thank Shane Mueller and Christoph Weidemann for comments and for providing us with the data from their experiment. Correspondence concerning this article should be addressed to J. D. Balakrishnan, Statistics Department, California Polytechnic State University, San Luis Obispo, CA 93407 (e-mail: jbalakri@calpoly.edu).

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. (2006). Objective analysis of classification behavior: Applications to scaling. In E. N. Dzhafarov & H. Colonius (Eds.), *Measurement and representation of sensations* (pp. 159-201). Mahwah, NJ: Erlbaum.

BALAKRISHNAN, J. D., MACDONALD, J. A., & KOHEN, H. S. (2003). Is the area measure a historical anomaly? *Canadian Journal of Experimental Psychology*, **57**, 238-256.

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

KATZ, L. (1970). A comparison of Type II operating characteristics derived from confidence ratings and from latencies. *Perception & Psychophysics*, **8**, 65-68.

KORNBROT, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, **68**, 393-414.

LINK, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Erlbaum.

LINK, S. W., & HEATH, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, **40**, 77-105.

LUCE, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

MUELLER, S. T., & WEIDEMANN, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, **15**, 465-494.

PARDUCCI, A., & SANDUSKY, A. (1965). Distribution and sequence effects in judgment. *Journal of Experimental Psychology*, **69**, 450-459.

RATCLIFF, R., & SMITH, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, **111**, 333-367.

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.

WEIDEMANN, C. T., & MUELLER, S. T. (2008). Decision noise may mask criterion shifts: Reply to Balakrishnan and MacDonald (2008). *Psychonomic Bulletin & Review*, **15**, 1031-1034.

NOTES

1. For our purposes, the distinction between continuous and discrete distributions (and, hence, density vs. relative frequency or area under a curve vs. height of a curve) is not important.

2. The sample sizes of these estimated likelihood ratios in each individual condition were relatively small (because the lowest confidence responses were relatively infrequent). However, combining the data across the six unequal base rate (9 to 1) conditions of the two experiments in Balakrishnan (1998a), there were 655 samples to estimate the value of $L(A_1)$, and the result was .93.

3. Because the sensory evidence is minimal when the participants' *yes*-or-*no* responses are extremely fast, the participants are, by definition, guessing (or at least virtually guessing) on these trials. In this particular data set, there is no analogous slow guessing (except for the one estimated point mentioned earlier, the functions in Figure 2 do not reach the value 1 on the outside edges). This could merely be due, however, to the sample sizes of the experiment (the slow RT bins had to be quite wide in order to reach the required 300 samples per bin) and to the small size of the bias manipulation (2 to 1 instead of 9 to 1). In support of this, the RT likelihood ratios for the slowest and fastest *A* responses in the combined unequal base rate conditions from Balakrishnan (1998a) were .87 and .91, respectively.

(Manuscript received December 21, 2007;
revision accepted for publication April 13, 2008.)