Perception & Psychophysics
2006, 68 (3), 393-414

Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation

DIANA EUGENIE KORNBROT
University of Hertfordshire, Hatfield, England

New and old methods of analyzing two-choice experiments with confidence ratings are evaluated. These include the theory of signal detectability (TSD), Luce's choice theory, *nonparametric* techniques based on areas under receiver-operating characteristic (ROC) functions, and methods based on S' and Ω, proposed by Balakrishnan and his colleagues. New methods for assessing the bias of a complete ROC function are proposed, together with an additional area-based measure of response bias. Area measures of both sensitivity and bias proved the most consistent. Response bias for a full ROC function was larger than bias at the cut point and also provided additional information. Participants showed voluntary control of bias for all measures except Ω. Unequal variance versions of TSD and choice models gave similar fits to data, with the choice model closer to an equal variance version. Discrimination data from Balakrishnan (1999) formed the empirical test bed.

The signal detection framework has been a cornerstone of discrimination research for more than half a century. Its key feature is the provision of distinct measures for sensitivity and bias that underpin basic and applied research into sensory and decision-making processes. Within this general framework, there are two main approaches: model based and nonparametric. The model-based approach makes explicit and testable assumptions about the distribution of the sensory representations of stimuli. By contrast, the nonparametric approach, which might preferably be termed *distribution free*, makes no assumptions about the form of sensory distributions. The aim of the present study is to evaluate measures of sensitivity and bias from both approaches. These include long established model-based measures from the theory of signal detectability (TSD) and Luce's choice theory, some recent nonparametric measures suggested by Balakrishnan and his colleagues (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003), and nonparametric area-based measures, with those for response bias presented here for the first time.

The present work was prompted by the challenges posed by Balakrishnan and his colleagues (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003). They suggested that currently used TSD and choice measures of sensitivity and bias are fatally flawed. This is worrying because basic and applied researchers need to be able to choose appropriate measures of sensitivity and bias, secure in the knowledge that conclusions based on these measures are not flawed. These challenges are addressed empirically, using one of Balakrishnan's (1999) own very comprehensive data sets. In addition to evaluating existing measures of sensitivity and bias, we provide a new area-based measure of bias and new methods for assessing bias that summarize *all* data points, in those paradigms that use confidence ratings. We also assess the relative strengths of more and less constrained versions of Luce's choice theory and TSD.

This article has three main sections. The theory section describes the key concepts, equations, and criteria necessary for empirical evaluations. The analysis section applies the criteria to empirical data from 4 individual participants making two-choice visual discriminations with confidence ratings (Balakrishnan, 1999). The final section discusses theoretical and practical implications of these analyses.

THEORY, EQUATIONS, AND CRITERIA

This section will start with a brief overview of two-choice experimental paradigms and the general signal detection approach, including an explication of the term *nonparametric*. Then key equations of the model-based approaches will be provided. This will be followed by a description of distribution-free methods, including the new area-based measure of response bias. Then a new procedure for measuring the bias of complete receiver-operating characteristic (ROC) functions will be proposed. Finally, evaluation criteria for measures of sensitivity and bias and for model evaluation will be summarized.

Thanks to colleagues who read drafts and made insightful and helpful comments: Ken Gilhooly, Elena Kulinskaya, Duncan Luce, Rachel Msetfi, and David Wellsted. Thanks also to the anonymous referees for constructive comments on earlier drafts. Correspondence about this article should be addressed to D. E. Kornbrot, University of Hertfordshire, College Lane, Hatfield AL10 9AB, England (e-mail: d.e.kornbrot@herts.ac.uk).

The most common experimental paradigm for discrimination in the signal detection framework is a simple two-choice experiment with two possible stimuli (a, b) and two possible responses (A, B). For example, the stimuli might be a shorter line (stimulus a) and a longer line (stimulus b) with responses of *short* (A) or *long* (B), as in the data analyzed here. This paradigm is also relevant to applied situations, where the stimuli might be healthy or diseased biopsy samples, with responses being *healthy* or *diseased* (see Macmillan & Creelman, 1991, for an excellent review). The A response to stimulus a is termed a *hit*, and the A response to stimulus b is termed a *false alarm*. The probability of a hit (h) and the probability of a false alarm (f) constitute the raw data from which sensitivity and bias measures are constructed, whether model based or nonparametric. In the confidence-rating version of this paradigm, participants give their chosen A or B response, together with a rating of their confidence in the accuracy of the response, given as c, on a scale from 1 to a maximum value of 100 (C_{MAX}). These values are combined by recoding the responses as follows. The A response with confidence c is coded with a negative value, as decision criterion $k = -c$, whereas the B response with confidence c is coded with a positive value, as decision criterion $k = +c$. The decision criteria (ks) thus range from -100 (A, very confident) to -1 (A, very unsure) and from $+1$ (B, very unsure) to $+100$ (B, very confident). All equations and figures are displayed in terms of the decision criterion k. Then, at each criterion k, one can calculate the values h_k (the probability of a hit, given criterion k) and f_k (the probability of a false alarm, given criterion k). This is done by labeling all responses less than or equal to k (A_k) and all responses greater than k (B_k) and calculating $h_k = p$ (response $\text{A}_k|$stimulus a) and $f_k = p$ (response $\text{B}_k|$stimulus a). There are thus 2 C_{MAX} (h_k,f_k) pairs. These may be used to generate an ROC function by plotting h_k as a function of f_k. Figure 1 shows examples of experimental raw ROC functions.

For the model-based approach (choice or TSD) one point sensitivity and one point bias measure may be calculated for each (h_k, f_k) pair. The exact equations depend on the model. The area-based approach also provides measures of sensitivity and bias for each (h_k, f_k) pair. The theory section will show how to calculate all these sensitivity and bias measures. The confidence-rating version of the two-choice paradigm may also be used to generate ROC measures of sensitivity and bias on the basis of several (h_k, f_k) pairs. Here, such measures are termed *ROC measures*, and detailed equations for the different approaches will be provided in the theory section.

Sensitivity measures are useful because they describe how effectively a particular participant performs a specified discrimination. A key feature of a sensitivity measure is that it is invariant with respect to changes in motivation due to changes in reward structure or a priori stimulus probability. It is posited to depend only on the current ability of the participant and the difficulty of the discrimination task. The invariance property of sensitivity measures can be tested by manipulating motivation, while holding person and stimuli constant. A very large number of per-

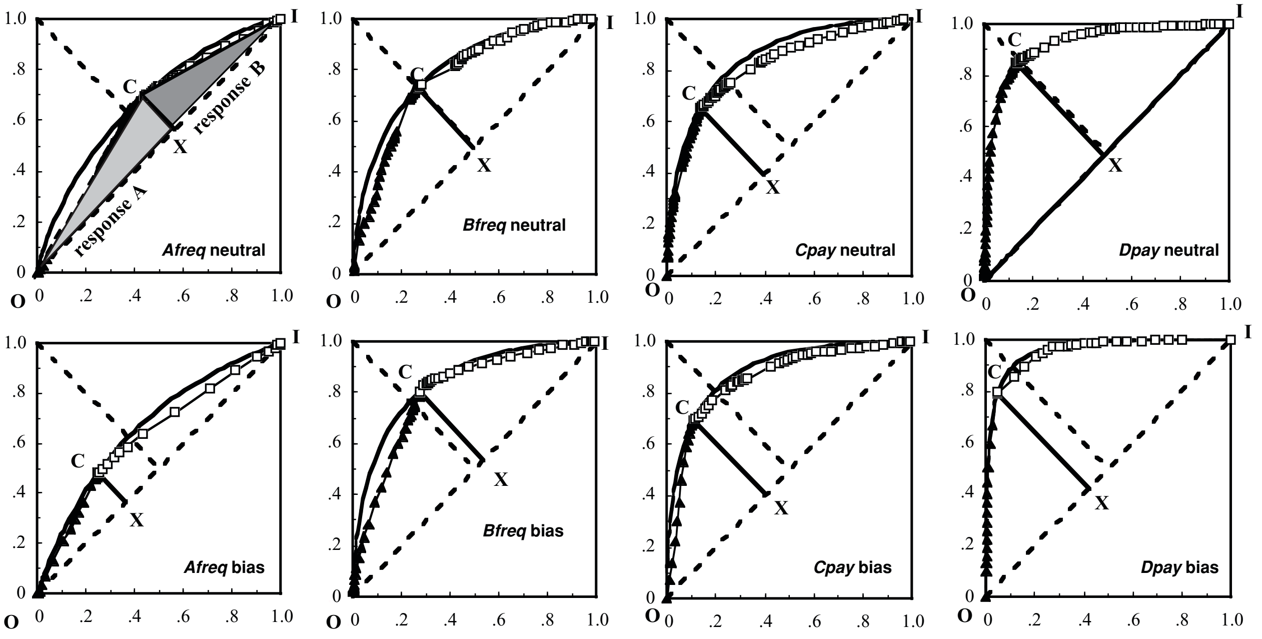

Figure 1. ROC function p(hit) (h) as a function of p(false alarm) (f) for all the participants. Top panels, neutral conditions; bottom panels, biased conditions. Filled triangles are from response A, open squares from response B. The top left panel shows the key areas shaded, as an example, with ΔOCX = K_A (corresponding to response A) and ΔICX = K_B (corresponding to response B). The point C represents the cut point between responses A and B. The line CX divides the region where response A is made from the region where response B is made.

ceptual and memory discrimination experiments using the separate condition paradigm have shown that model-based sensitivity measures do not change with bias condition (Macmillan & Creelman, 1991). By contrast, sensitivity measures derived from the rating paradigm generally do show some dependence on confidence rating.

A key feature of a bias measure is that it is posited to be under voluntary control but consistent across motivational conditions. Motivational factors comprise rewards for being right, punishments for being wrong, the relative a priori probability of stimuli, and pressures toward speed or accuracy. Another desirable property of a bias measure is that it should provide a measure of normatively optimal performance for any combination of a priori stimulus probability and payoffs. For any participant, a given payoff matrix and a priori stimulus probabilities should generate the *same* value of the bias parameter. This property can be fully tested only by tracing out an isobias function—that is, holding payoff constant and manipulating sensitivity. There are relatively few such studies (Dusoir, 1975, 1983; Irwin, Hautus, & Francis, 2001; Kornbrot, Galanter, & Donnelly, 1981; McCarthy & Davison, 1981, 1984). Most of the studies show rational conservatism: People move their decision criteria in the normatively correct direction, but less than would be predicted by normative models.

In general, response bias parameters are more easily interpretable if *symmetric* about a predicted value of *zero* for a *neutral* condition. However, traditional measures of response bias, based on likelihood ratios (often denoted β), have a value of 1 in a neutral condition, a value between 1 and infinity when biased toward the A response, and a value between 0 and 1 when biased toward the B response. All the bias measures, defined in the theory section as β values, have this property, including: β_{T} for TSD (denoted β_{G} by Macmillan & Creelman, 1991), β_{L} for choice theory, and the newly defined area bias measure β_K. Such β values are *not* symmetric about the neutral value of unity. For example, if $\beta = 1$ is neutral, favoring A twice as much as B would give $\beta = 2$, which is a difference of $2 - 1 = 1$ from neutral, whereas favoring B twice as much as A gives $\beta = 0.5$, which is a difference of $0.5 - 1 = -0.5$. This is *asymmetric* and gives the false impression that favoring A twice as much as B is further away from neutral than is favoring B twice as much as A. However, if one uses $\ln(\beta)$ as a measure of bias, the neutral bias gives $\ln(\beta) = \ln(1) = 0$; favoring A twice as much as B gives $\ln(\beta) = \ln(2) = +0.69$, whereas favoring B twice as much as A gives $\ln(\beta) = \ln(0.5) = -0.69$. Thus, the advantage of any measure $\ln(\beta)$ is that it is symmetric about the neutral point, so its magnitude is less likely to produce false impressions.

All approaches within the signal detection framework share the assumption that the internal representation of a stimulus over many trials generates an internal distribution on an internal variable, X (Macmillan, 2002; Macmillan & Creelman, 1991). The mean of the stimulus a distribution is assumed to be displaced from the mean of the stimulus b distribution by a distance, d. A participant is assumed to set a cut point or criterion, c, on the X dimension. On each trial, if the internal sensory representation is greater than c, response B is given; otherwise, response A is given. The representation on dimension X depends on external properties of the stimulus and internal attributes of the participant but is *not* under voluntary control. By contrast, the location of the criterion (c) is assumed to be under voluntary control. More controversially, participants are also assumed to *attempt* to set their criteria *optimally*, so as to maximize rewards and minimize penalties.

Approaches in the general signal detection framework may be divided into two broad classes, often designated *parametric* and *nonparametric*. The parametric class makes specific assumptions about the form of the stimulus and criterion representation distributions. It includes TSD (normal stimulus representation distribution) and Luce's choice model (logistic stimulus representation distribution). The nonparametric class makes no such distributional assumptions. Because of the confusions surrounding the parametric/nonparametric distinction, in this article parametric approaches are termed *model based*, whereas nonparametric approaches are termed *distribution free* (Macmillan, 2002; Macmillan & Creelman, 1991). In order to understand the reasons for this terminology, it is desirable to specify what exactly is meant by *nonparametric* in the signal detection context. The term *nonparametric* is generally applied to statistical procedures if either one or both of the following hold: (1) The variables are nonmetric, either ordinal or nominal, and/or (2) the distribution of the variables in the population is unknown. If only the second holds, the procedures may be more accurately termed *distribution free*. Ordinal statistical procedures—that is, procedures based on ranks—are also often termed *nonparametric*. However, ordinal procedures are usually *not* distribution free when applied to metric (interval or ratio) data. Common procedures in this class, including the Mann–Whitney, Wilcoxon, and Kruskal–Wallis, assume that all relevant distributions are of the *same* shape—that is, have the same variance, skew, kurtosis, and all higher moments. However, no assumption is made about what that shape actually is. In the signal detection framework, the variables at issue are the representations, assumed to be metric, of the physical stimuli and criteria in the human brain. Both TSD and choice models have sensitivity and bias measures that are distribution dependent. So-called *nonparametric* sensitivity and bias measures make no such distributional assumptions and, so, are distribution free. This includes both classic measures, such as the area under the ROC curve, and newer measures proposed by Balakrishnan and his co-workers. Nevertheless, even the distribution-free measures are metric parameters derived from the probabilities of hits and false alarms. Hence, the contrast of *model based* versus *distribution free* is preferred to the contrast of *parametric* versus *nonparametric*. An important corollary is that it makes sense to compare arithmetic means of a sensitivity or bias parameter between groups.

Model-Based Approaches

In general terms, the sensitivity measure for either TSD or Luce's choice model can be expressed as the distance, d, between the mean of the stimulus a distribution and the mean of the stimulus b distribution, divided by some estimate of variance from the stimulus b and stimulus a distributions. (Note that sensitivity in this signal detection sense is quite different from sensitivity of tests in the medical diagnostic sense.) If variances of the sensory distributions for the two stimuli differ (perhaps because stronger stimuli are more variable), there will be an additional sensory measure to describe the ratio of the variance of the stimulus a distribution to the variance of the stimulus b distribution. A measure of bias is chosen that is some function of the cut point c that would remain constant for different sensitivities if—and it is a big *if*—participants set c to maximize their objective rewards.

Theory of signal detectability. The simplest version of TSD (Macmillan & Creelman, 1991; Swets, 1986) has a sensitivity measure (d') and a bias measure (β_T; subscript "T" for TSD) that may be calculated from the proportion of hits (h) and the proportion of false alarms (f). The d' measure is defined as the separation between the mean of the stimulus b normal distribution and the mean of the stimulus a normal distribution, divided by their assumed common standard deviation (arbitrarily set to unity):

$$d' = z(h) - z(f), \tag{1}$$

where $z(p)$ is the inverse normal probability corresponding to cumulative probability (p) (Macmillan & Creelman, 1991, Equation 2.10).

The likelihood ratio bias parameter (β_T) is the ratio of the probability density (height of the curve) of the stimulus b distribution to the probability density of the stimulus a distribution at the cut point criterion, and $\ln(\beta_T)$ (Macmillan & Creelman, 1991, Equation 2.10) is the TSD measure that is symmetric about a neutral value of zero:

$$\ln\left(\beta_T\right) = .5\left[z(h)^2 - z(f)^2\right]. \tag{2}$$

Optimal values for β_T and hence, equivalently, $\ln(\beta_T)$ when there are biasing manipulations due to different a priori stimulus probabilities or payoffs are also easily obtained:

$$\beta_{\text{optimal}} = \left(\frac{\pi_a}{\pi_b}\right)\left(\frac{\text{payoff}(\text{A} \mid a) - \text{payoff}(\text{A} \mid b)}{\text{payoff}(\text{B} \mid b) - \text{payoff}(\text{B} \mid a)}\right), \tag{3}$$

where π_a, π_b are a priori probabilities of stimuli a and b, respectively.

When one can obtain several points on an ROC curve, either from a rating experiment or from several conditions with different optimal biases, more information is available. Then TSD predicts that z(correct | stimulus x) will be a linear function of z(error | stimulus x):

$$z(h) = +d'_{\text{ROC}} + \left(\frac{1}{s_T}\right)z(f), \text{ for response A,} \tag{4A}$$

and

$$z(m) = -d'_{\text{ROC}} + \left(\frac{1}{s_T}\right)z(cr), \text{ for response B,} \tag{4B}$$

where d'_{ROC} is sensitivity at the mean of the b distribution; $s_T = [\sigma(b)/\sigma(a)]$, the ratio of the stimulus b variance to the stimulus a variance; cr is probability of *correct reject*, p(response B | stimulus b); and m is probability of a *miss*, p(response B | stimulus a). Equations 4A and 4B represent the TSD normal transformed ROC functions. Estimates of the measures s_T and d'_{ROC} can be obtained directly from the slopes and intercepts of Equations 4A and 4B.

Obviously, given d'_{ROC} and s_T, one can calculate d' and $\ln(\beta_T)$ corresponding to any other empirically determined value of $z(f)$. Furthermore, one may calculate whether β_T is optimal for any experimenter-determined bias condition. If the ROC curve arises from separate experiments for each point, the ROC obtained from the response B from Equation 4B is completely determined by the conventional response A form, because $h + m = 1$ and $c + f = 1$. However, in a confidence-rating experiment, there are independent measures for response A and response B for each confidence level. In fact, the measure d'_e, suggested by Egan (Egan, Schulman, & Greenberg, 1959), is more comparable to d' at a neutral cut point; the parameter is the value of d' where the TSD ROC line cuts the minor diagonal and is given by (Macmillan & Creelman, 1991, Equation 3.8):

$$d'_e = 2d'_{\text{ROC}}\left(\frac{s_T}{1 + s_T}\right). \tag{5}$$

Equations 4A and 4B describe the TSD ROC in normal–normal coordinates and enable graphical and statistical evaluation of the TSD model. One may obtain estimates of s_T and d'_{ROC} (and hence, d'_e) as the mean of the values obtained from the response A and response B versions of the normal transformed ROC functions in Equations 4A and 4B. One may then test whether s_T = unity, thus supporting the simpler equal variance version of TSD. Obviously, with unequal variances, d' estimates at an arbitrary cut point are not predicted to be bias free. Note also that the stimulus variances (σ_B, σ_A) actually include the criterion (confidence-rating) k variances. This is because what is measured is the difference between the relevant stimulus mean and a criterion, and the variance of a difference var($x-y$) is the sum of var(x) + var(y). That is, z_h is generated from the cumulative distribution of the representation of the difference between the a stimulus and the criterion, and similarly for z_f. In a simple two-choice experiment, there is only one criterion and, so, only one criterion variance. Consequently, stimulus and criterion variance cannot be disentangled, and there is no point in considering the criterion variance separately. Rating experiments are more complex. There are several criteria and so, potentially, several different criterion variances. Equations 4A and 4B include the *implicit* assumption that all criterion variances are equal. If the variances of the extreme confidence ratings are higher than those for the less extreme confidence ratings, there will be systematic deviations from the linearity predictions of Equations 4A and 4B. If all criterion variances are equal, the whole ROC function can be predicted with just two free parameters, d'_{e} and s_{T}. These same arguments also apply to the choice formulation.

Luce's choice theory. Choice theory (Luce, 1959) can be cast in a form very similar to that of TSD. The only difference is that the logistic distribution is substituted for the normal distribution. For a single experiment, the sensitivity parameter equivalent to d' is $\ln(\eta)$, and the bias parameter is $\ln(\beta_{\mathrm{L}}$; subscript $_{\mathrm{L}}$ for Luce), where for any probability p, the logit of p, $\mathrm{lgt}(p)$, is given by

$$\mathrm{lgt}(p) = \ln\left[\frac{p}{(1-p)}\right]. \tag{6}$$

The choice sensitivity parameter, $\ln(\eta)$, is then given by (Macmillan & Creelman, 1991, Equation 2.13):

$$\begin{aligned} \ln(\eta) &= 0.5[\mathrm{lgt}(h) - \mathrm{lgt}(f)] \\ &= 0.5[\ln(h) - \ln(f) \\ &\quad + \ln(1-f) - \ln(1-h)]. \end{aligned} \tag{7}$$

The bias measure, β_{L}, is given by

$$\begin{aligned} \beta_{\mathrm{L}} &= h(1-h)/[f(1-f)] \\ \ln\left(\beta_{\mathrm{L}}\right) &= \ln(h) + \ln(1-h) - \ln(f) - \ln(1-f). \end{aligned} \tag{8}$$

In choice theory, $\ln(\beta_{\mathrm{L}})$ plays the same role as $\ln(\beta_{\mathrm{T}})$ in TSD. The optimal value of the criterion is thus also given by Equation 3.

The choice model is also convenient because asymptotic standard errors (*ASE*s) of the measures are simple to calculate. If some measure X is given by the ratio or product of two independent probabilities, the *ASE* of $\ln(X)$ is given by

$$ASE[\ln(X)] = \sqrt{\left[\frac{1}{n_1} + \frac{1}{N_1 - n_1} + \frac{1}{n_2} + \frac{1}{N_2 - n_2}\right]},$$

where for $i = 1, 2$, $p_i = n_i/N_i$ are the independent probabilities given by n_i criterion events from N_i attempts (Agresti, 1996). Application of this result gives the *ASE* for the choice parameters:

$$\begin{aligned} ASE[2\ln(\eta)] &= ASE\left[\ln\left(\beta_L\right)\right] \\ &= \sqrt{\left[\frac{1}{n_{\mathrm{A}a}} + \frac{1}{n_{\mathrm{A}b}} + \frac{1}{n_{\mathrm{B}a}} + \frac{1}{n_{\mathrm{B}b}}\right]}, \end{aligned} \tag{9}$$

where n_{Ji} is the number of responses "J" (J = A, B) to stimulus i ($i = a, b$). The equations for the choice ROC, equivalent to Equations 4A and 4B for TSD, are

$$\mathrm{lgt}(h) = +2\,\mathrm{lgt}(\eta)_{\mathrm{ROC}} + \left(\frac{1}{s_{\mathrm{L}}}\right)\mathrm{lgt}(f), \quad \text{for response A,} \tag{10A}$$

$$\mathrm{lgt}(m) = -\mathrm{lgt}(\eta)_{\mathrm{ROC}} + \left(\frac{1}{s_{\mathrm{L}}}\right)\mathrm{lgt}(cr), \quad \text{for response B,} \tag{10B}$$

where s_{L} = ratio of variances for a and b choice theory (stimulus − criterion) representations. Equations 10A and 10B represent a logistic or choice model ROC function. As with TSD, one may obtain an average estimate of the slope as the geometric mean of the slopes from the A and B responses as $\mathrm{lgt}(\eta)_{\mathrm{ROC}}$ and then obtain an estimate of the choice parameter at the cut point as $\mathrm{lgt}(\eta)_{\mathrm{e}}$, where

$$\mathrm{lgt}(\eta)_{\mathrm{e}} = \mathrm{lgt}(\eta)_{\mathrm{ROC}}\left[\frac{s_{\mathrm{L}}}{\left(1 + s_{\mathrm{L}}\right)}\right]. \tag{11}$$

It is often claimed that the logistic and normal distributions are so similar that one cannot distinguish Equations 4A and 4B from Equations 10A and 10B. However, there are differences for extreme ratings, because the normal density distribution drops very sharply as $\exp(-1/x^2)$, whereas the logistic density distribution drops only as $\exp(-1/x)$. Another difference is the ratio of variances, s_{T} or s_{L}. A variance ratio of unity indicates a simpler model with one less parameter. For auditory categorization of loudness, Kornbrot (1978, 1980, 1984) found variance ratios much closer to 1 for the normal model than for the logistic distribution.

Distribution-Free Approaches

Distribution-free measures of sensitivity attempt to estimate how "far away" the observed h, f pairs of probabilities are from values corresponding to no discrimination at all (the major diagonal in Figure 1), without making any assumptions as to the distribution of the sensory representation.

Area-based measures. A widely used distribution-free measure of sensitivity of this type is the area under the raw ROC function (i.e., h as a function of f), denoted here A. If only one point is available, an *estimate* of A (A') is given by (Macmillan & Creelman, 1991, Equation 4.8)

$$A' = 1 - \frac{1}{4\left(\dfrac{f}{h} + \dfrac{1-h}{1-f}\right)}. \tag{12}$$

This is actually the average of the minimum and maximum possible areas, given the observed values of f and h. (Craig, 1979; Macmillan & Creelman, 1996; Pollack & Norman, 1964). Clearly, A' is a *point* measure that can be calculated at each confidence-rating criterion. The extent to which A' changes with criterion is then an empirical question. The extent to which A' at any point is a good estimate of A, the area under the full ROC curve, is also an empirical question.

There have been several suggestions of area bias measures, but all have been shown to be monotonic with one of the choice model bias parameters and, hence, not distri-

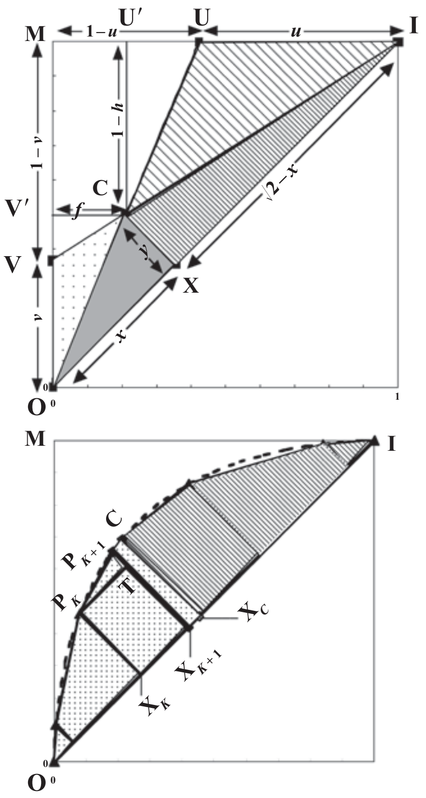

Figure 2. ROC plot of h as a function of f, showing areas required to calculate point area measures A', K'_A, and K'_B (top panel) and full ROC area measures A, K_A, and K_B (bottom panel). Dotted areas represent response A; striped areas represent response B. In the top panel, OCX is the minimum area for response A, OCX + OCV is the maximum area for response A, ICX is the minimum area for response B, and ICX + ICU is the maximum area for response B. In the bottom panel, the polygon $P_kP_{k+1}X_{k+1}X_k$ represents a generic polygon contributing to a total area K_A or K_B.

bution free (Craig, 1979; Grier, 1971; Hodos, 1970; Macmillan & Creelman, 1991, 1996). For this reason, they will not be discussed further. The area bias measure proposed here (β_K) is defined as the ratio K_B/K_A, where K_B is the area between the ROC curve and the major diagonal to the right of the minor diagonal (shown striped in Figure 2) and K_A is the area between the ROC curve and the major diagonal to the left of the minor diagonal (shown dotted in Figure 2). In the same spirit as Equation 12, one may then define *estimated* measures K'_A and K'_B as the averages of the minimum and maximum possible areas between the ROC curve and the major diagonals to the left and right of the minor diagonals. In the following equations, empirical areas are indicated by unprimed symbols (A, K_A, and K_B), whereas *estimates* that are based on the mean of the maximum and minimum possible areas, assuming a concave ROC function, are indicated by corresponding primed symbols (A', K'_A, K'_B, and β'_K). Obviously, $A = K_A + K_B + 0.5$, and $A' = K'_A + K'_B + 0.5$. Simple geometry (see the Appendix) gives the following equations for K'_A and K'_B:

$$K'_A = \frac{1}{4}(h-f)\left(h+f+\frac{f}{(1-f)}\right); \qquad (13A)$$

$$K'_B = \frac{1}{4}(h-f)\left(2-(h+f)+\frac{(1-h)}{h}\right). \qquad (13B)$$

An approximate area-based point bias measure is then given by

$$\beta'_K = \frac{K'_A}{K'_B} = \frac{\left(h+f+\dfrac{f}{(1-f)}\right)}{\left(2-(h+f)+\dfrac{h}{(1-h)}\right)}. \qquad (14)$$

The $\ln(\beta'_K)$ can be calculated at any criterion value, k, and has the desirable property of being symmetric about zero, where the criterion indices k run from $-C_{\text{MAX}}$ through 0 to $+C_{\text{MAX}}$. Neutral bias gives $\ln(\beta'_K)$ equals zero, and equivalent biases toward A and B responses give equal and opposite values of $\ln(\beta'_K)$. Consequently, $\ln(\beta'_K)$ can be directly compared with the model-based measures $\ln(\beta'_T)$ from TSD and $\ln(\beta_L)$ from the choice model. Furthermore, $\ln(\beta'_K)$ is not monotonic with any of the choice parameters.

When several points on an ROC curve are available, the empirical estimates of the ROC area measures K_A and K_B can be obtained. These can then be used to calculate the ROC area sensitivity measure, $A_{\text{ROC}} = 0.5 + K_A + K_B$ and the ROC area bias measure, $\beta_K = K_A/K_B$. If participants can give confidence ratings between 1 and C_{MAX}, the rating function has $2C_{\text{MAX}}$ points indexed by the criterion k, ranging from 1 to $2C_{\text{MAX}}$. Equations for K_A and K_B are given below (their derivation from simple geometry is given in the Appendix):

$$K_A = \frac{1}{4}\sum_{k=1}^{k=C_{\text{MAX}}}\left[\left(h_{k+1}-f_k\right)^2-\left(h_k-f_{k+1}\right)^2\right]; \qquad (15A)$$

$$K_B = \frac{1}{4}\sum_{k=C_{\text{MAX}}+1}^{k=2C_{\text{MAX}}}\left[\left(h_{k+1}-f_k\right)^2-\left(h_k-f_{k+1}\right)^2\right]. \qquad (15B)$$

Equations 15A and 15B can then be used to give a more accurate area bias measure at the A, B cut point, denoted β_K, analogous to the measure in Equations 14A and 14B:

$$\beta_K = \frac{K_A}{K_B}. \qquad (16)$$

The β_K measures behavior at the A, B cut point but uses performance from the whole ROC.

Note that K_A and K_B are areas fully under the ROC function, not approximations to a smooth curve going through a very large number of points. Consequently K_A and K_B will depend on the number of criteria. With a confidence rating scale from 1 to 100, the difference is negligible. However, with a small number of criteria (e.g., just confident and unconfident, giving only four possible responses), the underestimation would be substantial. Hence, A_{ROC} and β_K should be used only to compare conditions with the same number of responses. More accurate estimates of the average of the minimum and maximum possible areas are available from the author. However, they are tedious to calculate and show little or no advantage over K'_A and K'_B.

Balakrishnan's distribution-free measures for full ROC function. To obtain a sensitivity measure, Balakrishnan (1998b) considered the discrete function $U_R(j)$ defined over J possible rating criteria, defined as

$$U_R(k) = p(\text{response} \leq k \,|\, \text{stimulus } a) - p(\text{response} \leq k \,|\, \text{stimulus } b).$$

$U_R(k)$ is the difference between the hit and the false alarm probabilities at criterion k, which is, of course, the same as the difference between the probabilities of a correct and an erroneous response at criterion k. Balakrishnan (1998b) defined the sensitivity measure, S' as the sum of $U_R(k)$ over all criteria:

$$S' = \sum_{k=1}^{2C_{\text{MAX}}} U_R(k). \tag{17}$$

When the number of criteria is small—as, for example, with a Likert scale, where people rate responses as *uncertain*, *moderately confident*, or *very confident*—most (but not necessarily all) participants use all the responses. However, when the confidence is on a rating scale from 1 to 100 or on a continuous slider, people rarely use all the possible responses, and the S' measure is confounded by the *number* of different responses a person chooses to use. This measure is provided here because it was presented as an alternative to d' by Balakrishnan and his co-workers (Balakrishnan, 1998a, 1998b, 1999). He and his co-workers now suggest a further measure, γ_0, that estimates unbiased performance (Balakrishnan, MacDonald, & Kohen, 2003). Whether γ_0 remains constant across conditions with different a priori stimulus frequencies and/or different payoffs does not appear to have been tested. Hence, γ_0 will not be considered here.

Balakrishnan and his co-workers (Balakrishnan, 1998a, 1998b, 1999) also noted that prior to their work, all measures of bias were what we have called here *point* measures. Such measures in no sense assess the bias of a whole ROC function, relative to neutral. This is an important insight. It suggests that a useful bias measure should both assess the shift of the whole ROC function *and* be independent of the distribution of the stimulus representation. With these goals in mind, they define *suboptimal*—in their terms, *biased*—responses as responses where $p(\text{correct} \,|\, R = k) < 0.5$. For such a criterion, k, an estimate of bias ω_k is equal to the total number of responses at that value of k divided by the total number of trials (Balakrishnan, 1998b, note to Table 5).

Then,

$$\omega_{k\text{A}} = 0 \qquad \text{if } p(\text{A} \,|\, a) > p(\text{A} \,|\, b),$$
$$\omega_{k\text{A}} = \left[\frac{n(\text{A} \,|\, a) + n(\text{A} \,|\, b)}{N}\right] \quad \text{if } p(\text{A} \,|\, a) < p(\text{A} \,|\, b),$$
$$\Omega_\text{A} = -\sum_{k=1}^{k=C_{\text{MAX}}} \omega_{k\text{A}}; \tag{18A}$$

$$\omega_{k\text{B}} = 0 \qquad \text{if } p(\text{B} \,|\, b) > p(\text{B} \,|\, a),$$
$$\omega_\text{B} = \left[\frac{n(\text{B} \,|\, b) + n(\text{B} \,|\, a)}{N}\right] \quad \text{if } p(\text{B} \,|\, b) < p(\text{B} \,|\, a),$$
$$\Omega_\text{B} = +\sum_{k=C_{\text{MAX}}+1}^{k=2C_{\text{MAX}}} \omega_{k\text{B}}, \tag{18B}$$

where N is the total number of trials, Ω_A is a measure of bias on occasions when response A is given and is negative, and Ω_B is a measure of bias on occasions when response B is given and is positive. The total bias of the complete ROC function is then defined as $\Omega_\text{T} = \Omega_\text{A} + \Omega_\text{B}$. Positive values of Ω_T indicate bias toward B, and vice versa. Balakrishnan and his colleagues' approach to suboptimality and bias is thus different from that of the general signal detection framework. Their approach *first* identifies adjacent criteria k that are suboptimal according to the criterion $p(\text{correct}) < p(\text{error})$ and *then* uses ω_k as a measure of the magnitude of that bias. The ω_k are *point* measures based on simple probabilities, unlike either the area- or the model-based bias measures, in Equations 8, 14, and 16, which are point-based bias measures based on cumulative probabilities. Thus Ω_T is different from other bias measures both because it is a whole ROC measure and because it is ultimately based on simple, rather than cumulative, probabilities. So it is to be expected that Ω_T will behave differently from other bias measures.

The ln(odds ratio) as a distribution-free bias measure. A widely used distribution-free point measure of bias is the ln(odds ratio). It may be obtained by assuming that no matter how the independent probabilities h and f are generated, they remain constant throughout an experiment and, hence, both of their sample estimates are generated by a binomial distribution. The ln(odds ratio) is identical with the choice bias measure $\ln(\beta_\text{L})$. Clearly, the ln(odds ratio) may be calculated at any or all criterion values.

New Measures of Bias From Rating ROCs

A key property for a bias measure is that it should be used consistently for any given set of motivational factors. If the experimental situation is neutral—that is, stimuli have equal a priori probabilities and equal rewards for correct responses and penalties for errors—bias as measured by $\ln(\beta)$ for confidence c, given response A, should be the exact opposite of bias as measured by $\ln(\beta)$ for confidence c, given response B. If the conditions are biased, the function should be displaced. The general prediction is

$$\ln(\beta_{\text{Bmeasure}}) = -(\beta_{\text{Ameasure}}) + G_{\text{measure}}, \tag{19}$$

where *measure* can be derived from TSD, choice, or area procedures.

The constant should be zero for neutral conditions and negative for bias toward response B. The G_{measure} constants are *new* measures indicating how far a whole ROC function departs from being neutral, rather than how much the cut point between A and B is displaced from neutral. Equation 19 can be tested for β_T from Equation 2, β_L from

Equation 8, and β'_K from Equation 14. (One might also use β_K from Equation 16, but it is so similar to β'_K that a separate test is not worthwhile.) For model-based approaches, the optimal value of the constant is predictable from the payoff matrix and the a priori probabilities of the stimuli (see Equation 3). Equation 19 is important because it provides a way to evaluate the bias of the whole ROC function that depends on cumulative probabilities, unlike Ω, which depends on simple probabilities.

A slope of -1 is implicitly predicted for the regression of $\ln(\beta_{\text{Bmeasure}})$ on $\ln(\beta_{\text{Ameasure}})$ in Equation 19. If the slope is -1, there is symmetry about the G_{measure} point, and one can state that the participant is consistent in the usage of criteria but biased. Whether the bias is *appropriate* or optimal is then an empirical question that depends on the payoffs and a priori stimulus probabilities. A slope different from -1 indicates that a participant has a different confidence scale for A and B responses.

Evaluation Criteria

The evaluation criteria address three issues: performance of the sensitivity measures, performance of the bias measures, and assessment of fit of equal and unequal variance versions of TSD and choice models.

Performance of the sensitivity measures. The first evaluation criterion for the sensitivity measures is invariance of the point measures as a function of criterion, k, within each condition ($k = -C_{\text{MAX}}, -C_{\text{MAX}}+1, \ldots, -1, 0, +1, \ldots, C_{\text{MAX}}-1, C_{\text{MAX}}$). This will be evaluated in three ways. The first is visual inspection of a plot of relative sensitivity as a function of criterion k. Relative sensitivity, S_{rel}, is defined by

$$S_{\text{rel}} = \frac{(\text{sensitivity at criterion } k)}{(\text{sensitivity at cut point})}. \tag{20}$$

Visual inspection is obviously subjective and, so, is used only to determine whether further analysis is useful. Where two models are so similar that visual inspection is uncertain, any differences are unlikely to be important anyway, even if they are statistically significant. The second criterion is the frequency of occurrence of values of sensitivity more than an arbitrary x% from the value at the cut point. Analyses shown here use the criterion of $x\% = 15\%$, but essentially the same conclusion would have been drawn with $x\% = 5\%$ or 10%. The third criterion is the minimum and maximum percentages of overestimation and underestimation of sensitivity, relative to the cut point sensitivity, where

$$\text{overestimate \%} = \frac{100(\text{sensitivity at criterion } k)}{(\text{sensitivity at cut point}) - 100} \tag{21A}$$

and

$$\text{underestimate \%} = \frac{100 - 100(\text{sensitivity at criterion } k)}{(\text{sensitivity at cut point})} \tag{21B}$$

for over- and underestimates, respectively. The number of extreme criterion points included will affect all these criteria. Furthermore, both the number and the extremity of criterion use changes quite a lot across conditions and participants. For this reason, the variance (or coefficient of variation) of sensitivity measures is not a useful measure of consistency across criteria. These methods of evaluation may not be ideal, but they are explicit, so that other investigators can apply them to their own data—or suggest better ones.

It is also important to know whether the rating procedure has substantial advantages over the simpler two-choice procedure. To evaluate this issue, point measures will be compared with their ROC equivalents for TSD, choice, and area formulations.

Finally, the invariance of sensitivity measures across different bias conditions will be tested. If people and stimuli are unchanged across conditions, there should be no change in sensitivity measures. However, neutral conditions were always run first, so there might be practice effects.

Performance of the bias measures. The first criterion is the consistency of use of point bias measures across responses A and B. This criterion may be evaluated by testing Equation 19 for linearity and unit negative slope, separately for TSD, choice, and area measures.

Then performance of point bias measures at the cut point will be compared with equivalent ROC measures from Equation 19.

The next issue is whether participants have control of bias. This is evaluated by testing whether the bias measures have *different* values in *different* motivational conditions. Once it has been established that a bias measure does indeed change, the next question is whether the actual values of bias measures are *optimal*. In its strongest form, this question makes sense only for model-based approaches, since it is necessary to know the form of the stimulus representation to determine the optimal value according to equation. In a weaker form, one may ask whether the bias is in the "right" direction—that is, is response B made more often if a correct response to stimulus b is more highly rewarded or if stimulus b has a higher a priori probability? However participants may not behave optimally according to any measure. Although it is still of interest to discover whether performance is in fact optimal according to each specific bias measure and, if not, whether bias is at least in the normatively correct direction.

The "fit" of equal and unequal variance versions of TSD and Luce's choice model. This criterion will be evaluated via the TSD and choice ROC functions by testing the fit of Equations 4A and 4B (TSD) and Equations 10A and 10B (choice), separately for each participant in each condition. The more general models are satisfied if there is a strong linear trend and no significant higher order polynomial trends. In addition, there should be no significant differences in parameter estimates obtained from the A and B responses. Finally, slopes not significantly different from unity indicate that the simpler, equal variance version of a model is tenable.

ANALYSIS OF BALAKRISHNAN'S (1999) DATA

The data analyzed here have been described by Balakrishnan (1999) and are available online (Balakrishnan & MacDonald, 2003). They consist of individual data on 4 participants performing a difficult line length discrimination task, with confidence ratings. All 4 participants performed in both a neutral and a biased condition. Experiment 2 was a frequency manipulation experiment. There were 2 participants, identified here as *Afreq* and *Bfreq*. Both performed in a neutral (equal) condition, with equal frequencies for stimulus *a* and stimulus *b*, and a biased (unequal) condition in which the frequency of stimulus *b* was three times the frequency of stimulus *a*. Experiment 3 was a payoff manipulation experiment. There were 2 participants, identified here as *Cpay* and *Dpay*. Both performed in a neutral condition, with payoffs that did not depend on the stimulus presented, and a biased payoff condition in which the rewards for the correct responses and the penalties for the wrong responses to stimulus *b* were three times the rewards and penalties for equivalent responses to stimulus *a*. Optimal bias toward A was 1/3, in both the frequency and the payoff biased conditions. The neutral condition came first in both experiments. Allowed confidence ratings were 1–100, derived from a slider scale. All analyses were based on cumulative probabilities where the absolute frequency (numerator) was at least 5 and the number of stimulus presentations (denominator) was at least 640.

General Description of ROC Functions

Figure 1 shows ROC functions for all the participants in both conditions. The solid line parallel to the minor diagonal shows the transition from response A to response B. The theoretical TSD functions for the value of d' at the cut point are shown as the continuous curves through OCI in the eight panels of Figure 1.

The following features are evident in Figure 1. The empirical ROC function shows a performance lower than that predicted by TSD (choice would be effectively identical). Participants *Bfreq* and *Dpay* appear very close to unbiased in the neutral condition. Participants *Afreq*, *Cpay*, and *Dpay* are all appropriately biased toward response B in the biased condition. However *Cpay* was just as biased toward response B in the neutral condition. Thus, 3 out of 4 participants did show voluntary control of bias. Two out of these 3, *Afreq and Dpay*, appear to have changed their bias in the *optimal* direction.

For all measures of sensitivity or bias, when an estimate is given, followed by two numbers in parentheses, these numbers are the 95% confidence limits. Exceptions to this convention are explicitly noted.

Performance of Sensitivity Measures

Invariance of point sensitivity measures as a function of criterion. Figures 3 and 4 show d', $\ln(\eta)$, and A' at criterion k *relative* to their values at the response A to response B transition (as defined by Equation 20) as a function of criterion, k, for Experiment 2 (frequency manipulation) and Experiment 3 (payoff manipulation), respectively. Relative sensitivity is defined by Equation 20. The functions for relative A' appear flatter than the other functions. Table 1 summarizes the characteristics of the functions in Figures 3 and 4. For each participant and condition, the three measures d', $\ln(\eta)$, and A' are compared according to three criteria. The first numeric column gives the number of points evaluated. For each such point, the ratio of that point's sensitivity measure to the same sensitivity measure at the cut point was calculated. The next column in Table 1 gives a count of the number of points for which the deviation of this ratio from unity was more than 15%. This count was considerably smaller for A' than for the other measures, for all the participants except *Dpay*. For *Dpay*, the choice model had fewer deviations greater than 15%, but the number for A' was still small. As is evident in Figures 3 and 4, the sensitivity at the cut point was near the maximum for all the models. Furthermore, the percentages of deviations from cut point values were much larger for underestimates than for overestimates. The final two columns of Table 1 show these percentages of deviations separately for underestimates and overestimates, according to Equation 21. Here, the A' measure is numerically superior for all the participants in all the conditions. This superiority of A' was a surprise, since there is no theoretical reason why A' should be invariant with respect to confidence.

Comparison of point and ROC sensitivity measures. Figure 5 shows ROC sensitivity measure as a function of point measures for TSD, choice, and area approaches. For TSD and choice, the regressions are excellent fits, with adjusted r^2 greater than 99%, and form close to the identity relation (slope = 1, intercept = 0). The area ROC and point measures were slightly less similar [adjusted $r^2 = .971$, slope = 1.25 (1.05, 1.45), intercept = −0.24 (−0.40, 0.057)]. So overall, these data show almost identical performance for ROC and point sensitivity measures and, hence, no advantage for the more complex confidence-rating procedure.

Comparison of sensitivity measures between neutral and biased conditions. Table 2 shows sensitivity measures for each participant in two different conditions, one neutral and one biased. All the measures, except S', show slight superiority for the biased condition for every participant except *Afreq*. For the choice point sensitivity measure $\ln(\eta)$, and for the ROC measures d'_e and $\ln(\eta_e)$, where standard errors are available, these effects are statistically significant at the 95% confidence level. Thus, 3 people performed better in the biased condition, and 1 performed (very slightly) better in the neutral condition. Balakrishnan's S' shows a different pattern. It is larger,

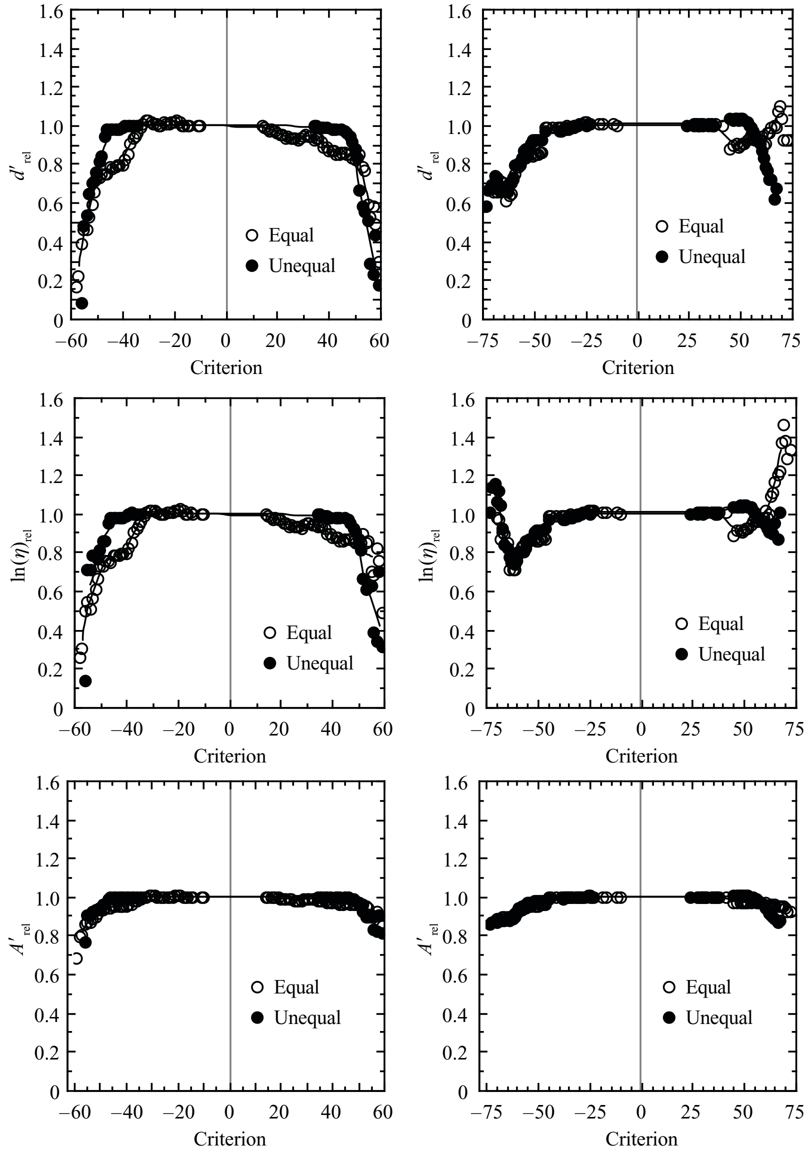

Figure 3. For the frequency manipulation participants, relative point sensitivity measures (as defined by Equation 20) at criterion *k*, relative to value at cut point between responses A and B, as a function of confidence rating. Left panels, *Afreq*; right panels, *Bfreq*. Top panels, relative *d′*; middle panels, relative ln(*η*); bottom panels, relative *A′*.

sometimes much larger, for the neutral condition for all the participants. The reason is simple: The participants used fewer points on the rating scale in the biased conditions, so there were fewer values of $(U_k - U_{k-1})$ to sum over (Table 1 gives number of points).

Performance of the Bias Measures

The behavior of $\ln(\beta_K)$ and $\ln(\beta'_K)$ are almost identical. For all 4 participants in both conditions, the adjusted r^2 values of regressions of $\ln(\beta_K)$ on $\ln(\beta'_K)$ were more than .99; slopes were not significantly different from 1, and intercepts were not significantly different from 0. Hence, only the more easily calculated $\ln(\beta'_K)$ will be used in most of what follows.

Consistency of point bias measures as a function of criterion across responses A and B. Figures 6 and 7 show the consistency of usage of the confidence level for the bias measures $\ln(\beta_T)$, $\ln(\beta_L)$, and $\ln(\beta'_K)$ by testing Equation 19. Table 3 shows intercepts = G_{measure}, slopes, and effect sizes (adjusted r^2 values) for the functions in Figures 6 and 7. As might be expected, TSD and choice give very similar results. For the most part, there is a consistent linear relation, with a high adjusted r^2. The exceptions are for *Afreq* in the biased condition for TSD and choice and

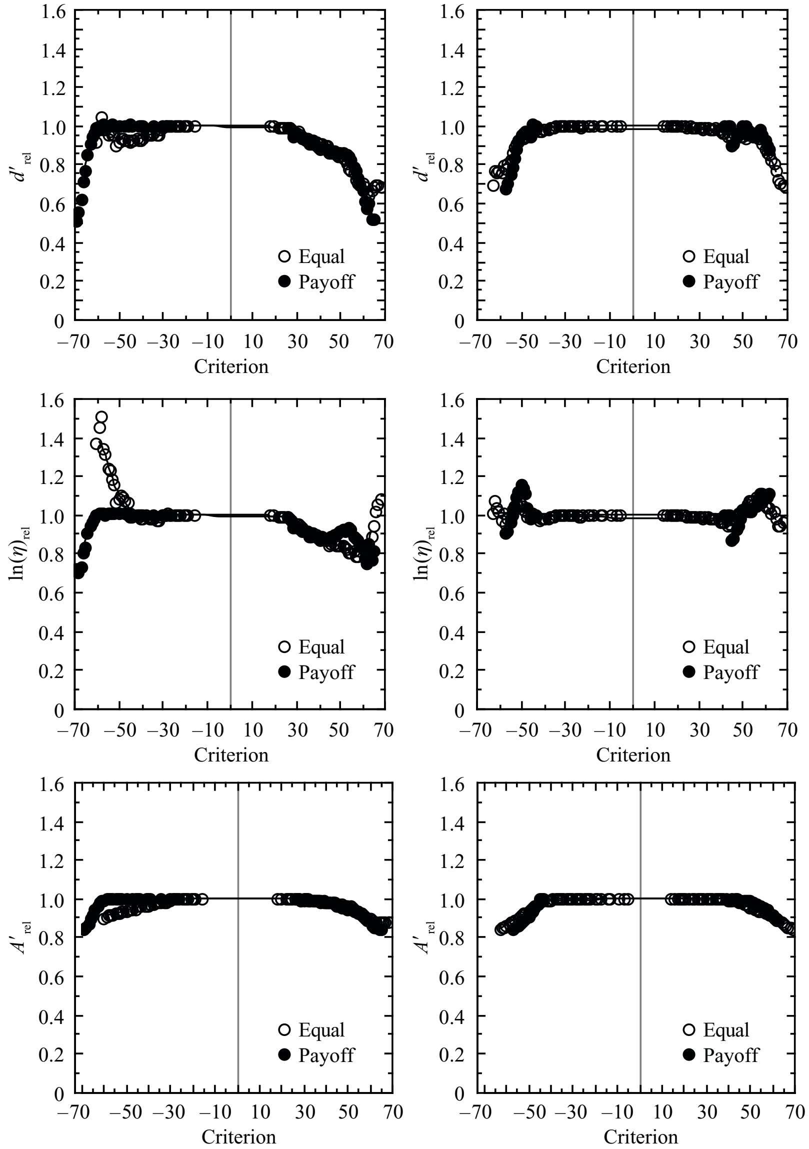

Figure 4. For the payoff manipulation participants, relative point sensitivity measures (as defined by Equation 20) at criterion *k*, relative to value at cut point between responses A and B, as a function of confidence rating. Left panels, *Cpay*; right panels, *Dpay*. Top panels, relative *d′*; middle panels, relative ln(*η*); bottom panels, relative *A′*.

for *Cpay* in the biased condition for all the models. This indicates that the use of any criterion, given a B response, can be predicted from the use of that same criterion, given an A response. Nevertheless, in most cases, the slopes are significantly different from 1, indicating that the subjective spacing between confidence criteria is systematically different for responses A and B. The poor fits for *Afreq* in the biased conditions may be due to the restricted range of bias used for both A and B responses, although the area function is an excellent fit (adjusted $r^2 = .997$). The poor fit for *Cpay* in the biased condition for all the analyses is due to the very restricted range for response A.

Comparison of point and ROC bias measures. Figure 8 shows ROC bias G_{measure}s as a function of their equivalent point measures for area, choice, and TSD formulations. Two features are apparent from Figure 8 and from the numeric values of the bias measures in Table 4. First, for all the participants and all the conditions, ROC measures are considerably (at least 1.6 times) larger than the equivalent point measures. Second, the behavior of *Cpay* in the biased condition is substantially different from that of the other 3 participants, as is also evident in Figure 7. If one excludes *Cpay* in the biased condition, regressions in Figure 7 all have intercepts not significantly

Table 1
Performance of Point Sensitivity Measures as a Function of Confidence Criteria

Participant	Condition	No. Criteria	Sensitivity	No. Outside 15%	Max % Overestimate	Max % Underestimate
Afreq	Neutral	97	TSD: *d′*	33	2.5	82.9
			Choice: ln(η)	29	2.4	74.6
			Area: *A′*	3	0.6	31.7
	Biased	60	TSD: *d′*	22	0.3	91.7
			Choice: ln(η)	20	0.4	86.1
			Area: *A′*	8	0.1	23.6
Bfreq	Neutral	101	TSD: *d′*	20	10.4	38.8
			Choice: ln(η)	16	46.1	29.1
			Area: *A′*	–	0.5	11.8
	Biased	79	TSD: *d′*	25	3.4	42.0
			Choice: ln(η)	11	4.2	25.9
			Area: *A′*	–	0.1	13.6
Cpay	Neutral	97	TSD: *d′*	29	4.9	36.0
			Choice: ln(η)	21	50.5	46.9
			Area: *A′*	–	0.1	12.4
	Biased	85	TSD: *d′*	18	0.5	48.2
			Choice: ln(η)	12	1.0	24.9
			Area: *A′*	3	0.1	16.1
Dpay	Neutral	119	TSD: *d′*	16	0.2	30.6
			Choice: ln(η)	–	11.2	4.6
			Area: *A′*	4	< 0.05	15.8
	Biased	38	TSD: *d′*	5	0.4	32.6
			Choice: ln(η)	1	16.0	13.1
			Area: *A′*	2	0.1	15.9

different from 0 at the 95% confidence level. The regression slopes are area = 1.64 (1.37, 1.92), TSD = 2.18 (1.84, 2.51), and choice = 2.13 (1.82, 2.45).

Thus, bias as estimated by an entire ROC is generally greater than bias estimated at the A, B cut point. It is also evident that for some of the participants, the behavior estimated from the ROC bias measure is different from the behavior estimated from the point bias measure. In particular, *Cpay* shows a very large ROC criterion shift (much more than the other participants), together with a rather small point criterion shift (less than the other participants).

Comparison of neutral and biased conditions: Voluntary control of bias measures. Table 4 shows *point* measures of bias—ln(β_{T}), ln(β_{L}), ln(β_K'), and ln(β_K)—at the cut point between responses A and B, together with ROC G_{measure}s from Equation 19 and Balakrishnan's Ω for each participant in a neutral and a biased condition. All of the measures except Ω show *different* behavior in the biased and the neutral conditions and, hence, voluntary control. For the point measures, statistical tests of these differences are available for the choice model, using the *ASE*s in Equation 9, and all differences are statistically significant at the 95% confidence level. For the G_{measure}s, standard errors are available from the regressions in Equation 19. Again, neutral and biased conditions show significant differences for all the participants.

Optimality of the bias measures. Table 4 shows little support for any strong version of optimality. The only per-

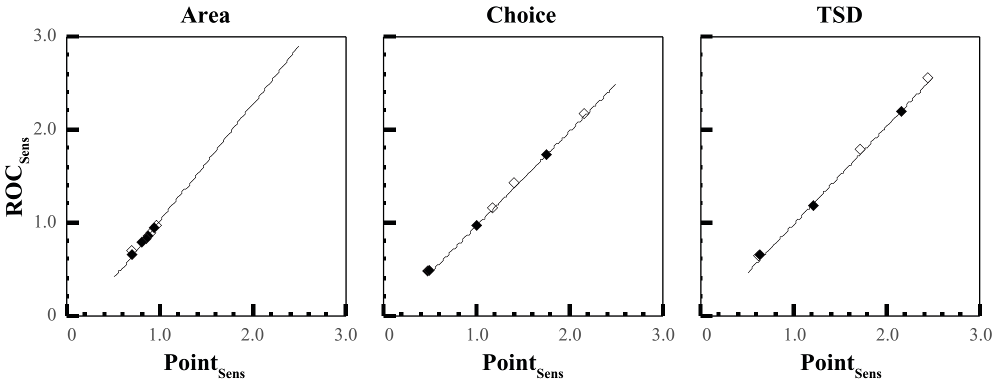

Figure 5. ROC sensitivity measures as a function of point sensitivity measures at the A, B cut point for the area, choice, and theory of signal detectability (TSD) approaches for the frequency manipulation participants in the neutral condition (filled symbols) and biased conditions (open symbols).

Table 2
Point and ROC Sensitivity Measures in Neutral and Biased Conditions

		Point		ROC	
Analysis	Participant	Neutral	Bias	Neutral	Bias
		d'		d'_e	
TSD	*Afreq*	0.67	0.63	0.65	0.63
	Bfreq	1.24	1.45	1.21	1.43
	Cpay	1.49	1.73	1.41	1.80
	Dpay	2.15	2.49	2.16	2.59
		$\ln(\eta)$		$\ln(\eta_e)$	
Choice	*Afreq*	0.54	0.52	0.51	0.51
	Bfreq	1.01	1.18	0.96	1.14
	Cpay	1.24	1.45	1.10	1.46
	Dpay	1.81	2.18	1.79	2.18
		A'		A	
Area	*Afreq*	.71	.70	.64	.67
	Bfreq	.82	.85	.78	.81
	Cpay	.85	.87	.82	.86
	Dpay	.92	.93	.92	.95
				Ω	
Balakrishnan	*Afreq*			20.1	8.8
	Bfreq			37.6	32.4
	Cpay			37.8	31.8
	Dpay			70.4	20.9

son with a bias parameter not significantly different from 0 in the neutral conditions is *Bfreq*. The only measure not significantly different from the normatively optimal value of -1.10 in a biased condition is $\ln(\beta_T)$ for *Dpay*.

By contrast, the weaker proposition that participants move their criteria in the normatively correct direction has considerable support from all of the measures except Ω. Table 4 shows the value of measures in the biased condition minus their equivalent values in the neutral condition. A negative value significantly different from zero, indicating a normatively correct move in bias toward B, is present for participants *Afreq*, *Cpay*, and *Dpay* for all ROC measures. *Afreq*, *Cpay*, and *Dpay* also show a statistically significant move in the expected direction for the choice point measure $\ln(\beta_L)$ at the A, B cut point (where a test is possible because the *ASE* is available). As was noted above, *Bfreq* moves in the normatively wrong direction on all measures. For point measures, the results for *Cpay* are equivocal, being small toward A for TSD and choice but small toward B for the area measure.

The values of Ω are very low for all the participants in all the conditions and, so, are uninformative. Furthermore, comparisons across conditions are not possible, since the standard error of Ω is not known.

Fit of TSD and Luce's Choice Model

Figures 9 and 10 show three versions of ROC function for Experiments 2 and 3, respectively. The choice model (bottom panels) appears to be a better fit than does TSD (middle panels), in the sense that slopes appear more similar across conditions and responses. This apparent superiority of the choice model is evaluated more rigorously by testing the regressions posited in Equations 4A and 4B and Equations 10A and 10B. The strong curvature apparent in most raw ROC functions rules out threshold models, which will not be discussed further.

The fit of TSD and choice model will first be evaluated by testing the linearity of their respective ROC functions separately for response A and response B, for all the participants in both conditions. There is little to choose between TSD and choice. Both models had an adjusted r^2 greater than .88 for all functions and greater than .98 for 15/16 functions (the poor fit was for *Dpay*, response A biased). Both models showed some nonlinear effects in terms of a quadratic component significant at the 99% confidence level for 6 out of 16 functions. The results gave small but significant differences in variance ratio (s_T or s_L) and/or sensitivity at the cut point [d'_e or $\ln(\eta_e)$] from the stimulus *a* and stimulus *b* versions of Equation 4 for all TSD functions except for *Dpay* in the neutral condition (Equations 4A and 4B) and for all choice functions except for *Dpay* in both the neutral and the biased conditions (Equations 10A and 10B). Thus, on the grounds of linear fit and differences between stimulus *a* and stimulus *b* estimates of variance ratios and sensitivity at the cut point, choice and TSD models give similar levels of fit.

The question of equal variance was evaluated by testing whether average estimates of s_T and s_L from response A and response B regressions (Equations 4A and 4B for TSD and Equations 10A and 10B for choice) were reliably different from unity. Table 5 shows estimates of s_T and s_L, together with their standard errors (s_T and s_L estimates significantly different from unity are shown in bold). Violations were tested at the 99% confidence level to ensure that the simpler equal variance model was not rejected without good cause. For TSD, all eight estimates

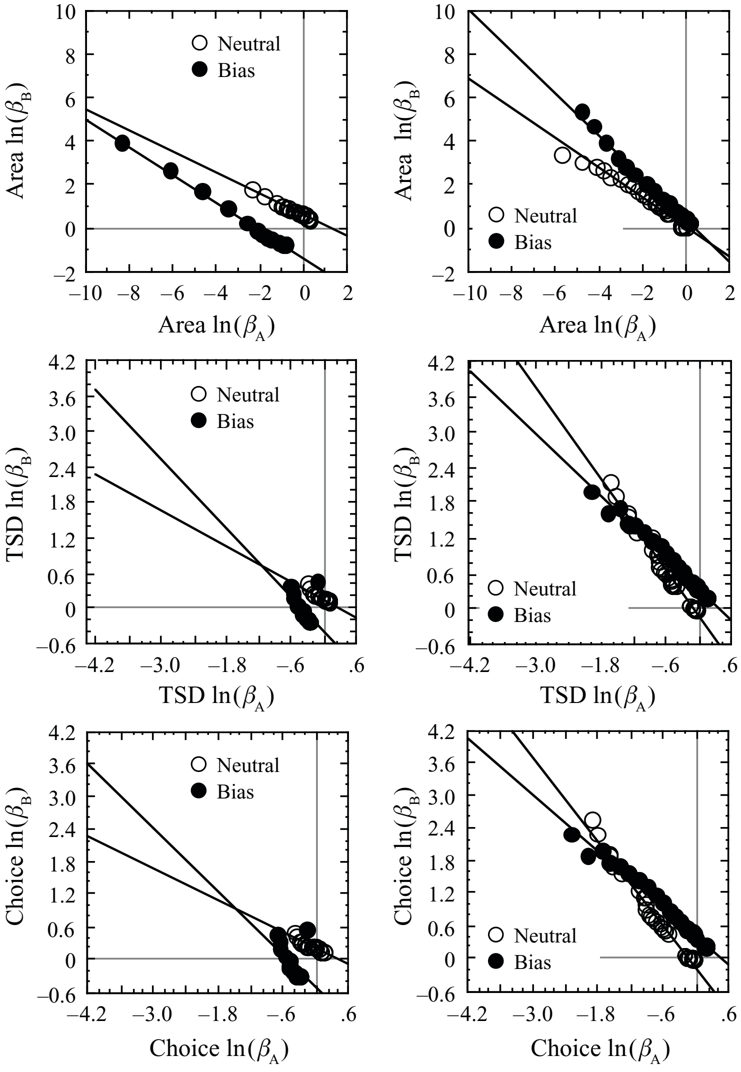

Figure 6. Response B bias measure as function of response A bias measure for the frequency manipulation participants. Top panels, area measure ln(β'_K); middle panels, theory of signal detectability (TSD) measure ln(β_T); bottom panels, choice measure ln(β_L).

of s_T are significantly less than unity (mean s_T = 0.79; 99% confidence limits, 0.72, 0.86). So the equal variance version of TSD is emphatically rejected. For the choice model, four estimates are not reliably different from unity, three are lower and one is higher than unity (mean s_L = 0.99; 99% confidence limits, 0.87, 1.10). Thus, the equal variance version of choice model is viable for some of the participants. The ANCOVA also show no evidence that slopes are different in biased and neutral conditions, for either TSD or choice.

DISCUSSION

The Signal Detection Approach

On the basis of the analyses presented here, as well as on the vast body of existing literature going back to the 1950s, the signal detection approach provides a useful framework for describing discrimination. A distribution-free version provides reliable measures of sensitivity, *A′*, and bias, ln(β'_K), for simple two-choice experiments, even without confidence ratings. Given rating data, one can compare models based on different distributions.

Sensitivity

Consistency. All the measures of point sensitivity showed an effect of confidence rating within conditions. Figures 3 and 4 show that there is a substantial middle range of confidence ratings where sensitivity remains constant and very similar to the value at the A, B cut point. Sensitivity values estimated from very extreme confidence ratings are lower than those at the cut point. Surprisingly, as is documented by the measures in Table 1, the *A′* mea-

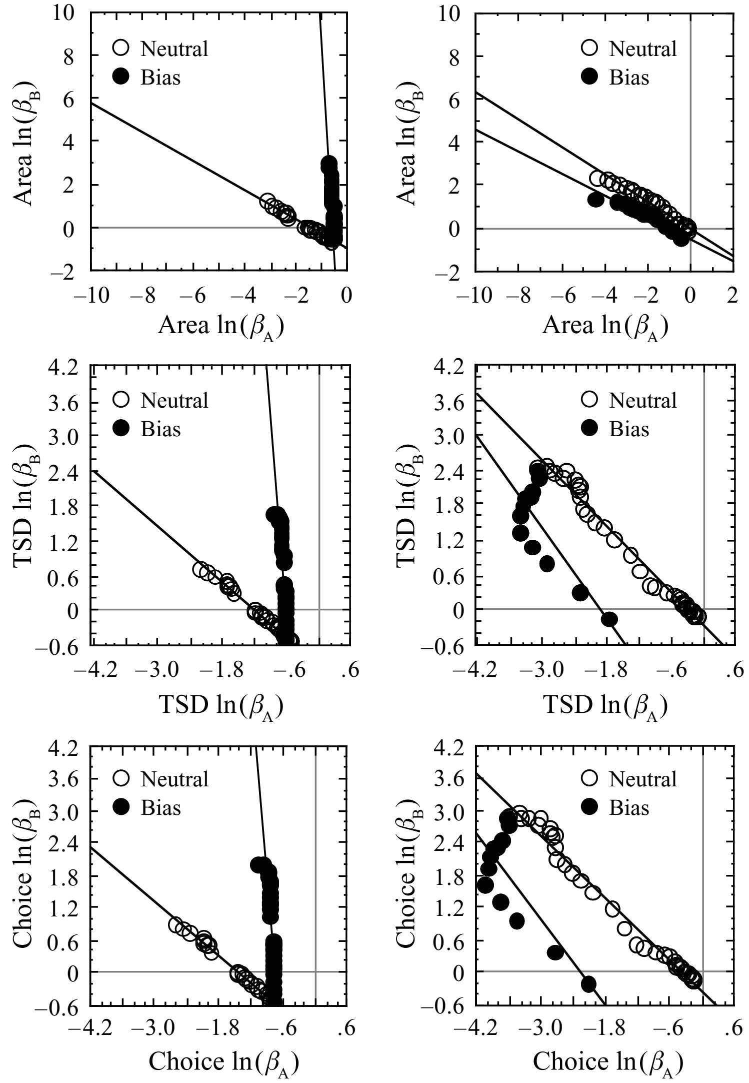

Figure 7. Response B bias measure as function of response A bias measure for the payoff manipulation participants. Top panels, area measure ln(β'_K); middle panels, theory of signal detectability (TSD) measure ln(β_T); bottom panels, choice measure ln(β_L).

sure shows the *least* variation over different confidence ratings. This *might* be an artifact of the fact that the range of *A′* is limited from .5 to 1.0, whereas ln(η) and *d′* range from 0 to unlimited.

Comparing point and ROC measures. The values of point sensitivity measures are very similar to the values of equivalent ROC measures. This is equally true for the area measure and the model-based measures. Consequently, in terms of accuracy of sensitivity measurement, confidence ratings provide no advantages. This is useful information. There may well be a tendency to take confidence ratings under the erroneous and time-consuming hypothesis that accuracy will be improved.

The ROC measure *S′* showed *lower* sensitivity in the biased condition for all the participants, due to differential use of criteria. In fact, consistent with the present analysis, Treisman (2002) has already proposed that unequal use of criteria would invalidate the use of *S′*. Balakrishnan has argued that such unequal use does not make much difference (Balakrishnan & MacDonald, 2002). The flaw in using *S′* is that of attempting to obtain areas from summing lines, with no width.

Comparing neutral and biased conditions. There were small but statistically reliable differences in sensitivity between the biased and the neutral conditions for the area and model-based measures. These favored the biased condition (performed second for all the participants) for 3 out of the 4 participants.

Bias

Consistency. Figures 6 and 7, based on Equation 19, provide a new way of measuring consistency of criterion

Table 3
Testing Equation 19: G_{measure} Slopes and Adjusted r^2 for Regressions of Bias Measure for Response B as a Function of Bias Measure for Response A Shown in Figures 6 and 7

		G_{measure}		Slope		Adjusted r^2	
Participant	Analysis	Neutral	Bias	Neutral	Bias	Neutral	Bias
Afreq	Area	0.58	−1.36	−0.48	−0.64	.991	.997
	Choice	0.22	−0.50	−0.50	−0.99	.895	.275
	TSD	0.17	−0.40	−0.50	−0.98	.895	.311
Bfreq	Area	0.02	0.36	−0.68	−0.98	.978	.994
	Choice	−0.14	0.42	−1.29	−0.87	.983	.987
	TSD	−0.11	0.34	−1.28	−0.88	.985	.988
Cpay	Area	−1.04	−10.67	−0.69	−19.15	.997	.776
	Choice	−1.19	−9.16	−0.84	−12.12	.994	.628
	TSD	−0.92	−7.50	−0.79	−12.31	.992	.641
Dpay	Area	−0.07	−0.55	−0.64	−0.51	.981	.929
	Choice	−0.31	−3.05	−0.96	−1.35	.989	.886
	TSD	−0.25	−2.45	−0.94	−1.29	.989	.893

use across different responses. Individual participants show substantial consistency as assessed by the adjusted r^2 values in Table 3. The area bias measure, $\ln(\beta'_K)$, appears to be slightly more consistent than the TSD bias measure, $\ln(\beta_T)$, or the choice bias measure, $\ln(\beta_L)$. The adjusted r^2 is highest for the area measure in seven out of eight comparisons. This is similar to the finding for sensitivity. The slopes of the functions in Figures 6 and 7 are *not* equal to −1. Thus, participants typically impose a different scale for confidence for A and B responses. This is also a new and far from obvious finding.

Comparing point and ROC performance. Figures 6 and 7 also provide ways of measuring the shift of an entire ROC function via the values of G_{measure}. Armed with this measure, one can compare point and ROC bias measures. Values of β'_K from Equation 14 and β_K from Equation 15 are very similar. So, ratings give no advantage in terms of accuracy of measurement of bias at the cut point.

By contrast, comparing G_{measure} for the entire ROC function with equivalent point measures gives a different picture. As is shown in Table 4, the ROC measures are always higher. Even the neutral conditions show some degree of bias, using G_{measure}. It thus appears that the full ROC is more sensitive to deviations form neutrality than are the point measures. Furthermore, considering the full ROC function provides information not available from point measures alone. The large change in behavior between neutral and biased conditions for *Cpay* is detectable only in Figure 7 and by the high values of G_{measure} in Table 3.

By contrast, Balakrishnan's Ω, also a full ROC measure, shows minimal bias for all the participants in all the conditions. It is not obvious what advantages there might be in a bias measure that does *not* actually change when people's decision making does show a change in bias on a raw—and hence, distribution free—ROC function (see Figure 1).

Comparison of neutral and biased conditions: Voluntary control and optimality. All the measures except Ω show voluntary control of bias, in that values of at least some bias measures are different in the neutral and

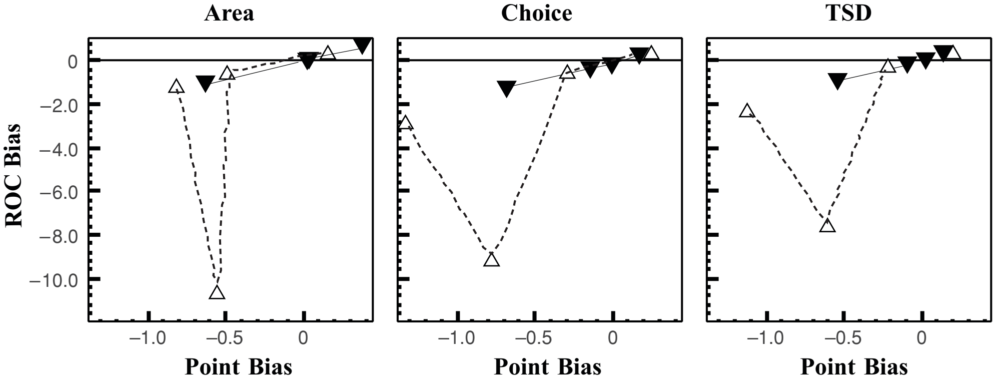

Figure 8. ROC bias measures as a function of point bias measures at the A, B cut point for area, choice, and theory of signal detectability (TSD) approaches for 4 participants in two conditions each. Biased conditions have up-pointing open triangles with dashed lines; neutral conditions have down-pointing filled triangles with solid lines. The outlier point represents participant *Cpay* in the biased condition.

Table 4
Point and ROC Bias Measures in Neutral and Biased Conditions

		Point at A, B cut			ROC at A, B cut			Full ROC from Equation 19		
Analysis	Participant	Neutral	Bias	B−N	Neutral	Bias	B−N	Neutral	Bias	B−N
			$\ln(\beta_T)$						G_{TSD}	
TSD	*Afreq*	0.12	−0.23	−0.35				0.17	−0.40	−0.57
	Bfreq	−0.01	0.17	0.18				−0.11	0.34	0.45
	Cpay	−0.53	−0.62	−0.09				−0.92	−7.50	−6.58
	Dpay	−0.12	−1.08	−0.96				−0.25	−2.45	−2.20
			$\ln(\beta_L)$						G_{choice}	
Choice	*Afreq*	0.15	−0.29	−0.44				0.22	−0.50	−0.72
	Bfreq	−0.01	0.22	0.23				−0.14	0.42	0.56
	Cpay	−0.66	−0.77	−0.11				−1.19	−9.16	−7.97
	Dpay	−0.14	−1.31	−1.17				−0.31	−3.05	−2.74
			$\ln(\beta'_{A'})$			$\ln(\beta_A)$			G_{area}	
Area	*Afreq*	0.38	−0.80	−1.18	0.24	−0.72	−0.97	0.58	−1.36	−1.93
	Bfreq	−0.02	0.14	0.16	−0.12	0.10	0.22	0.02	0.36	0.34
	Cpay	−0.62	−0.56	0.06	−0.57	−0.64	−0.07	−1.04	−10.67	−9.63
	Dpay	−0.07	−0.47	−0.40	−0.08	−0.54	−0.47	−0.07	−0.55	−0.48
									Ω	
Balakrishnan	*Afreq*							.06	.00	−.06
	Bfreq							.01	.02	.01
	Cpay							.04	.00	−.04
	Dpay							.00	.00	.00

Note—B−N is (value in biased condition) − (value in neutral condition). Negative values indicate a shift toward response B.

the biased conditions. In terms of point measures, *Afreq*, *Cpay* (minimally), and *Dpay* all change their bias in the normatively correct direction, whereas *Bfreq* changes in the opposite direction. This pattern is observed for all the measures. Shifts are larger for the full ROC than for the A, B cut point, as a necessary consequence of the finding that the deviation from neutrality is larger for $G_{measure}$s than for point measures.

Equation 3 predicts an optimal value $\ln(\beta_L)$ or $\ln(\beta_T)$ of −1.08 in the biased condition. There is little evidence to suggest that these participants chose this optimal bias. In terms of the raw ratio of proportion of A responses, relative to proportion of B responses, people with lower sensitivities need to be *more* biased. This did not happen, or did not happen sufficiently, so the least sensitive participant, *Afreq*, shows the weakest bias in terms of $\ln(\beta_L)$ or $\ln(\beta_T)$. It may be possible to train people to set their criteria optimally, but most (like these participants) are suboptimal without such training.

In real-world applications, different situations may have both different a priori probabilities and different payoffs. For example, malignant cells are less frequent in screening conditions than in biopsy conditions. The implications of errors are also different for different categories of response, such as *definitely malignant*, *possibly malignant*, *probably benign*, *benign*, and so forth. Similar arguments apply to the probability and degree of threat of different kinds of military weapons or computer viruses, or of different kinds of investments. Most laboratory rating experiments, like the one analyzed here, do not have clear predictions of optimality away from the cut point, because there is no greater penalty for being wrong about an extremely confident A response than an extremely tentative A response. Clearly, the behavior of bias measures as a function of criterion and motivation merits further exploration for both practical and theoretical reasons. It would seem that both point and ROC measures would be required.

Table 5
Tests of Equal Variance Versions of TSD and Choice Model

	Neutral		Bias	
Participant	Slope	*SE*(Slope)	Slope	*SE*(Slope)
		s_T From TSD		
Afreq	0.79	0.007	0.81	0.009
Bfreq	0.87	0.018	0.78	0.014
Cpay	0.89	0.012	0.64	0.013
Dpay	0.76	0.013	0.79	0.059
Mean	0.83		0.75	
LCL	0.77		0.69	
UCL	0.89		0.81	
		s_L From Choice		
Afreq	0.85	0.008	0.87	0.012
Bfreq	**1.08**	0.026	**0.99**	0.019
Cpay	**1.17**	0.021	0.81	0.013
Dpay	**1.03**	0.012	**1.09**	0.072
Mean	**1.03**		**0.94**	
LCL	0.95		0.86	
UCL	1.11		1.02	

Note—Slopes *not* significantly different from 1 at the 99% confidence level are in bold. Standard errors of slopes are obtained by averaging standard errors obtained from the stimulus *a* and stimulus *b* regressions. Lower confidence levels (LCLs) and upper confidence levels (UCLs) are based on standard errors that are averages of the eight values of SE^2 for the 4 participants in two conditions each.

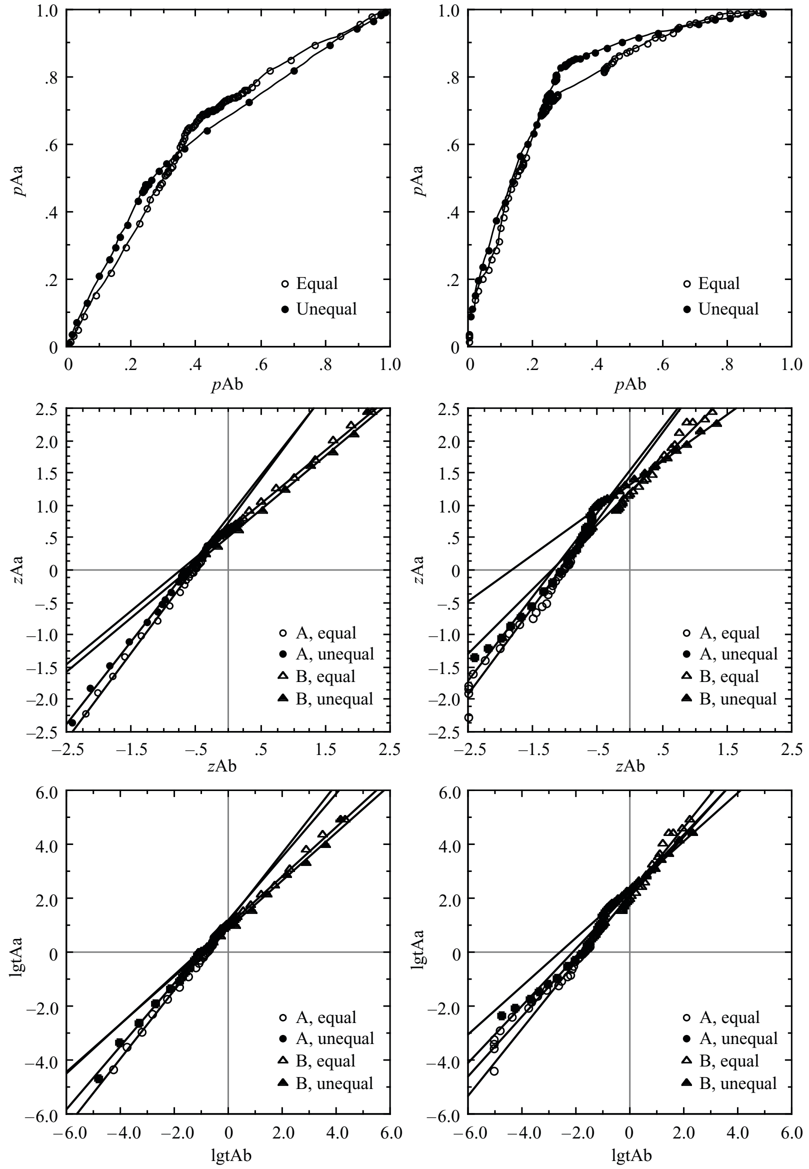

Figure 9. ROC functions for the frequency manipulation participants: *G*(*h*) as a function of *G*(*f*). Top panels, raw probabilities, *G* is identity transformation; middle panels, theory of signal detectability, *G* is normal transformation; bottom panels, Luce's choice model, *G* is logistic transformation.

Model Evaluation and ROC Functions

Balakrishnan (1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003) suggested that the empirical ROC functions do not fit any signal detection model. Figures 9 and 10 and Table 5 challenge this suggestion. Both TSD and choice model ROC functions are a "reasonable" fit to the model in terms of linear predictions of transformed ROC functions, with adjusted r^2 values generally greater than .98. The different variance ratios from stimulus *a* and stimulus *b* do not invalidate the models. Nevertheless, the fits of both TSD and the choice model do show systematic deviations from theory. Thus, neither the logistic nor the normal distribution provides an ideal representation of the effects of repeated presentations of the stimuli. The vindication of the signal detection approach arises from the finding of consistent estimates of sensitivity, dependent on people and stimuli, and consistent measures of bias, under voluntary control. Suboptimality of bias measures suggests further avenues for investigation, rather than a flaw in the approach.

For the line length discrimination task analyzed here, the equal variance version of the choice model cannot be rejected, because the choice variance ratio measure s_L is so close to unity. By contrast, the TSD variance ratio measure, .79, is substantially less than unity, suggesting that the stimulus *a* distribution has higher variance than does the stimulus *b* distribution. The choice model is to be preferred to

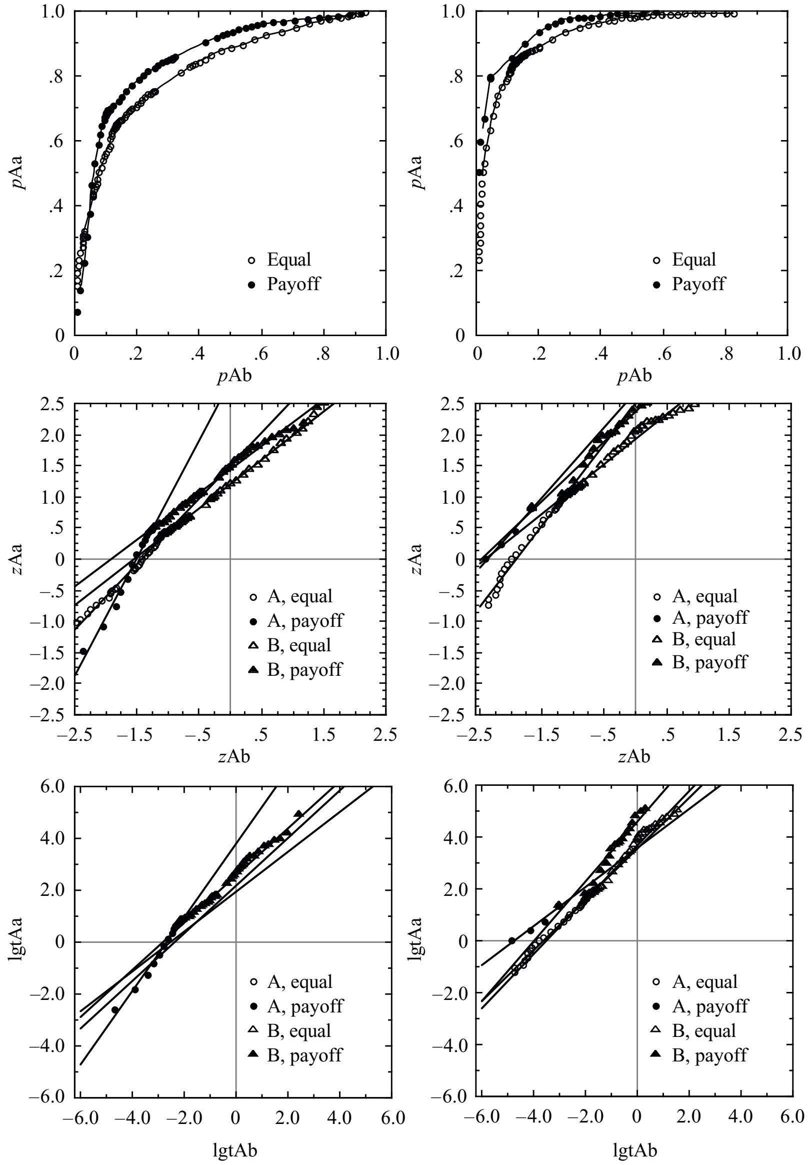

Figure 10. ROC functions for the payoff manipulation participants: *G*(*h*) as a function of *G*(*f*). Top panels, raw probabilities, *G* is identity transformation; middle panels, theory of signal detectability, *G* is normal transformation; bottom panels, Luce's choice model, *G* is logistic transformation.

TSD for this line length discrimination because the simpler version with a single sensitivity measure is acceptable.

Model-Based and Distribution-Free Approaches

The distribution-free approaches have the advantage of making fewer assumptions than do the model-based approaches. Furthermore, area-based measures, both old and new, have been shown to have clear advantages of robustness and consistency and are thus to be highly recommended. No advantages for either *S*′ or Ω emerge from these analyses.

Nevertheless, model-based approaches are clearly essential for deeper understanding of underlying processes. For example, TSD or choice measures of bias and sensitivity should be derivable from information accrual models, such as a version of the random walk model (Green & Luce, 1973; Heath & Fulham, 1988; Kornbrot, 1988; Laming, 1968, 1979; Link, 1975; Luce, 1986; Smith & Vickers, 1989; Stone & Callaway, 1964; Vickers, Caudrey, & Willson, 1971).

SUMMARY

The main findings may be summarized as follows.

1. The signal detection approach is successful and useful and not flawed, as was suggested by Balakrishnan and his colleagues.

2. Area measures are the best distribution-free measures. The new area bias measure, $\ln(\beta'_K)$, complements the well-established sensitivity measure A'.

3. Area measures are at least as good as TSD or choice measures for practical purposes.

4. Point measures of sensitivity and of bias at the cut point are just as good as ROC measures and much simpler to obtain.

5. The new techniques for assessing the bias of complete ROCs are important and give more information and information that is different from that given by bias at the cut point alone.

6. There are small but significant departures from the predictions of the choice model and TSD.

7. The simpler equal variance version of the choice model is acceptable, whereas TSD requires an extra parameter for the ratio of stimulus *a* to stimulus *b* variance.

The time-honored signal detection framework has been rigorously tested and emerged with flying colors. Choice theory is rather higher up the mast than is TSD.[1]

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis*. Chichester, U.K.: Wiley.

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.

Balakrishnan, J. D., & MacDonald, J. A. (2003). *Alternatives to signal detection theory*. Retrieved August 17, 2003, from www.psych.purdue.edu/~beowulf/dsdt/dsdt.html.

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.

Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Human Factors*, **21**, 69-78.

Dusoir, A. E. (1975). Treatments of bias in detection and recognition models: A review. *Perception & Psychophysics*, **17**, 167-178.

Dusoir, A. E. (1983). Isobias curves in some detection tasks. *Perception & Psychophysics*, **33**, 403-412.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America*, **31**, 768-773.

Green, D. M., & Luce, R. D. (1973). Speed–accuracy trade off in auditory detection. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 547-569). New York: Academic Press.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75**, 424-429.

Heath, R. A., & Fulham, R. (1988). An adaptive filter model for recognition memory. *British Journal of Mathematical & Statistical Psychology*, **41**, 119-144.

Hodos, W. (1970). A nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin*, **74**, 351-354.

Irwin, R. J., Hautus, M. J., & Francis, M. A. (2001). Indices of response bias in the *same–different* experiment. *Perception & Psychophysics*, **63**, 1091-1100.

Kornbrot, D. E. (1978). Theoretical and empirical comparison of Luce's choice model and logistic Thurstone model of categorical judgment. *Perception & Psychophysics*, **24**, 193-208.

Kornbrot, D. E. (1980). Attention bands. *British Journal of Mathematical & Statistical Psychology*, **33**, 1-16.

Kornbrot, D. E. (1984). Mechanisms for categorisation. *British Journal of Mathematical & Statistical Psychology*, **37**, 84-198.

Kornbrot, D. E. (1988). Random walk models of binary choice. *Acta Psychologica*, **69**, 109-127.

Kornbrot, D. E., Galanter, E. G., & Donnelly, M. (1981). Estimates of utility function parameters. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 441-458.

Laming, D. R. J. (1968). *Information theory of choice reaction times*. London: Academic Press.

Laming, D. R. J. (1979). A critical comparison of two random-walk models for two-choice reaction data. *Acta Psychologica*, **43**, 431-453.

Link, S. W. (1975). The relative judgment theory of two-choice reaction time. *Journal of Mathematical Psychology*, **12**, 114-135.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1986). *Response times*. Oxford: Oxford University Press, Clarendon Press.

Macmillan, N. A. (2002). Signal detection theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (pp. 43-90). New York: Wiley.

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

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, **3**, 164-170.

Macmillan, N. A., Rotello, C. M., & Miller, J. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics*, **66**, 406-421.

McCarthy, D., & Davison, M. (1981). Towards a behavioral theory of bias in signal detection. *Perception & Psychophysics*, **29**, 371-382.

McCarthy, D., & Davison, M. (1984). Isobias and alloiobias functions in animal psychophysics. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**, 390-409.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, **1**, 125-126.

Smith, P. L., & Vickers, D. (1989). Modelling evidence accumulation with partial loss in expanded judgment. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 797-815.

Stone, G. C., & Callaway, E. (1964). Effects of stimulus probability on reaction time in a number-naming task. *Quarterly Journal of Experimental Psychology*, **16**, 47-55.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks. *Psychological Bulletin*, **99**, 181-198.

Treisman, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review*, **9**, 845-857.

Vickers, D., Caudrey, D., & Willson, R. [J.] (1971). Discriminating between the frequency of occurrence of two alternative events. *Acta Psychologica*, **35**, 151-172.

NOTE

[1] Macmillan and his colleagues (Macmillan, Rotello, & Miller, 2004) provide methods for estimating the accuracy of several signal detection parameters. However, the author was unaware of these measures at the time of writing.

APPENDIX
Equations for Point and ROC Area Measures

Estimation of Areas K_A' and K_B'

The top panel of Figure 2 shows a single cut point, *C*, with coordinates *f*, *h*, on an ROC function, together with the triangles needed to estimate K_A' and K_B'.

Then, an estimate for K_A' is the minimum area bounded by the major diagonal and the cut line XC below and to the left of the ROC function (ΔOXC), plus *half* the ΔOVC that would need to be added to OXC to obtain the maximum area. This is the same pragmatic approach as that used by Pollack and Norman (1964; see also Macmillan & Creelman, 1996). So,

$$K_A' = \text{area } \Delta\text{OXC} + 0.5 \text{ area } \Delta\text{OVC}.$$

Similarly, an estimate for K_B' is the minimum area bounded by the major diagonal and the cut line XC below and to the right of the ROC function (ΔIXC), plus *half* the ΔIUC that would need to be added to IXC to obtain the maximum area; thus,

$$K_B' = \text{area } \Delta\text{IXC} + 0.5 \text{ area } \Delta\text{IUC}.$$

To calculate these areas, the distance *x*, *y*, *u*, and *v*, shown in the top panel of Figure 1, are needed. The values of *x* and *y* are the coordinates of the point *C*, when the *f*, *h* axes are rotated through 45º (using sin 45 = cos 45 = $1/\sqrt{2}$); so,

$$x = \frac{(h+f)}{\sqrt{2}},$$

and

$$y = \frac{(h-f)}{\sqrt{2}}.$$

The value of *u* may be obtained from the similar ΔOV′C, ΔOMU by noting that

$$\frac{1-u}{f} = \frac{1}{h};$$

therefore,

$$u = 1 - \frac{f}{h}.$$

Similarly, the value of *v* may be obtained from the similar ΔIU′C, ΔIMV by noting that

$$\frac{1-v}{1-h} = \frac{1}{1-f};$$

therefore,

$$v = 1 - (1-h)(1-f).$$

Using these values for the distances *x*, *y*, *u*, *v* gives the following equations for K_A', K_B':

$$\begin{aligned} K_A' &= \text{area } \Delta\text{XOC} + 0.5 \text{ area } \Delta\text{VOC} = \frac{xy}{2} + \frac{1}{2}\frac{vf}{2} \\ &= \frac{h^2-f^2}{4} + \frac{f(h-f)}{4(1-f)} \\ &= \frac{h-f}{4}\left(h+f+\frac{f}{1-f}\right) \qquad \text{(A1)} \end{aligned}$$

$$\begin{aligned} K_B' &= \text{area } \Delta\text{ICX} + 0.5 \text{ area } \Delta\text{ICU} = \frac{\left(\sqrt{2-x}\right)y}{2} + \frac{1}{2}\frac{u(1-h)}{2} \\ &= \frac{2(h-f)}{4} - \frac{h^2-f^2}{4} + \frac{(h-f)(1-h)}{4h} \\ &= \frac{h-f}{4}\left(2-(h+f)+\frac{1-h}{h}\right). \qquad \text{(A2)} \end{aligned}$$

APPENDIX (Continued)

Areas Under the Empirical ROC Curve, K_A, K_{Bx}

When a full ROC function is available, the actual areas to the left and right of the cut line CX_c may be obtained. If participants can give either of two responses and confidence ratings from 1 to C_{MAX}, there are $2C_{MAX}$ criteria, k from 1 to $2C_{MAX}$. This is illustrated in the bottom panel of Figure 2. K_A, dotted on the figure, is the area bounded by the empirical ROC function and the major diagonal and is to the left of the cut line CX_c. K_B, striped on the figure, is the area bounded by the empirical ROC function and the major diagonal and is to the right of the cut line CX_c. The kth point on the empirical ROC function, P_k, has coordinates f_k, h_k. Then, the areas K_A, K_B are obtained by summing polygons of the general form $X_kP_kP_{k+1}X_{k+1}$. The distance OX_k along the major diagonal is denoted x_k, and the distance X_kP_k parallel to the minor diagonal is denoted y_k. Then,

$$x_k = \frac{h_k + f_k}{\sqrt{2}};\ y_k = \frac{h_k - f_k}{\sqrt{2}}. \tag{A3}$$

The polygon $X_kP_kP_{k+1}X_{k+1}$ is composed of the rectangle $X_kP_kTX_{k+1}$, plus the triangle $P_kP_{k+1}T$, and so has an area given by

$$\begin{aligned}\text{area } X_kP_kP_{k+1}X_{k+1} &= y_k\left(x_{k+1} - x_k\right) + \frac{\left(y_{k+1} - y_k\right)\left(x_{k+1} - x_k\right)}{2}\\ &= \frac{\left(h_{k+1} - f_k\right)^2 - \left(h_k - f_{k+1}\right)^2}{2}.\end{aligned}$$

The area K_A is then given by summing all polygons from $k = 0$ to $k = c$, whereas the area K_B is given by summing all polygons from $c+1$ to 2CR. There are $2C_{MAX}$ criteria and, hence, $2C_{MAX}$ polygons. The point P_0 is the origin (0,0), and the point P_{2CR+1} is the point (1, 1). So,

$$K_A = \frac{1}{4}\sum_{k=1}^{k=c}\left[\left(h_{k+1} - f_k\right)^2 - \left(h_k - f_{k+1}\right)^2\right]$$

and

$$K_B = \frac{1}{4}\sum_{k=c+1}^{k=2CR+1}\left[\left(h_{k+1} - f_k\right)^2 - \left(h_k - f_{k+1}\right)^2\right]. \tag{A4}$$

(Manuscript received October 1, 2003;
revision accepted for publication May 19, 2005.)