

Running Head: P284 Distribution Free and Model Based Signal Detection

Signal Detection Theory, the Approach of Choice: Model Based and Distribution Free Measures  
and Evaluation

Diana Eugenie Kornbrot  
University of Hertfordshire

May 2005

Author Note

Professor Diana Eugenie Kornbrot, Psychology Department, University of Hertfordshire. s.

Thanks to Professor Balakrishnan who made the data analyzed here available on his web site, <http://www.psych.purdue.edu/~beowulf/dsdt/dsdt.html>, and also provided helpful emails and comments. Thanks to colleagues who read drafts and made insightful and helpful comments, Ken Gilhooly, Elena Kulinskaya, Duncan Luce, Rachel Msetfi, David Wellsted. Thanks also to anonymous referees and to Editor Prof. Maddox for constructive comments on earlier drafts.

Correspondence about this article should be addressed to: D. E. Kornbrot, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9AB, UK. +44 170 728 4626

[d.e.kornbrot@herts.ac.uk](mailto:d.e.kornbrot@herts.ac.uk)

### Abstract

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; "non-parametric" techniques based on *areas* under Receiver Operating Characteristic functions; and those based on  $S'$  and  $\Omega$ , proposed by Balakrishnan and his colleagues. New methods for assessing the bias of a complete Receiver Operating Characteristic 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  $\Omega$ . 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.

### Signal Detection Theory, the Approach of Choice

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 “non-parametric”. 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. This Ms. aims firstly to evaluate measures of sensitivity and bias from both approaches. These include: long established model based measures from TSD and Luce’s choice theory; some recent “non-parametric” measures suggested by Balakrishnan and his colleagues (Balakrishnan, 1998a; 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003a); and. “non-parametric” *area* based measures, with those for response bias presented here for the first time.

The current work was prompted by the challenges posed by Balakrishnan and his colleagues (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003a). They suggest 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 own very comprehensive data sets (Balakrishnan, 1999). 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.

The Ms. 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 four 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 starts with a brief overview of two choice experimental paradigms and the general signal detection approach, including an explication of the term “non-parametric”. Then key equations of the model based approaches are provided. This is 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 ROC functions is proposed. Finally evaluation criteria for measures of sensitivity and bias and for model evaluation are summarized.

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 “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 “healthy” or “diseased”. See Macmillan & Creelman, (1991), for an excellent review. The response “A” to stimulus  $a$  is termed a hit and the response “A” 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 “non-parametric”. 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. Response “A” with confidence  $c$  is coded with a negative value, as decision criterion  $k = -c$ , while the response “B” with confidence  $c$  is coded with a positive value, as decision criterion  $k = +c$ . The decision criteria  $k$ , thus range from  $-100$  (“A”, very confident) through  $-1$  (“A”, very unsure) and  $+1$  (“B”, very unsure) to  $+100$  (“B”, very confident). All equations and Figures are in displayed terms of the decision criteria  $k$ . Then at each criterion,  $k$ , one can calculate 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<sub>k</sub>” and all responses greater than  $k$  “B<sub>k</sub>” and calculating  $h_k = p(\text{response “A}_k” | \text{stimulus } a)$  and  $f_k = p(\text{response “B}_k” | \text{stimulus } a)$ . There are thus  $2C_{MAX}(h_k, f_k)$  pairs. These may

be used to generate a raw Response Operating Characteristic (ROC) function by plotting  $h_k$  as a function of  $f_k$ . Figure 1 shows examples of experimental raw ROC functions.

---

Insert Figure 1 about here

---

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 shows 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 based on several  $(h_k, f_k)$  pairs. Such measures are termed here *ROC* measures, and detailed equations for the different approaches are 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 perceptual and memory discrimination experiments from the separate condition paradigm show 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 towards 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 only be fully tested 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 studies show rational conservatism. People move their decision criteria in the normatively correct direction, but less than 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  $\beta$ ), have a value of 1 in a neutral condition, a value between 1 and infinity when biased towards response “A” and a value between 0 and 1 when biased towards response “B”. All the bias measures, defined in the theory section as  $\beta$  values, have this property: including:  $\beta_T$  for TSD (denoted  $\beta_G$  by McMillan and Creeleman, 1991),  $\beta_L$  for choice theory, and the newly defined *area* bias measure  $\beta_K$ . Such  $\beta$  values are *not* symmetric about the neutral value of unity. For example, if  $\beta = 1$  is ‘neutral’, then favoring “A” twice as much as “B” would give  $\beta = 2$ , which is a difference of  $2 - 1 = 1$  from neutral; while 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 favoring “B” twice as much as “A”. However, if one uses  $\ln(\beta)$  as a measure of bias, then the neutral bias gives  $\ln(\beta) = \ln(1) = 0$ ; favoring “A” twice as much as “B” gives  $\ln(\beta) = \ln(2) = +0.69$ ; while favoring “B” twice as much as “A” give  $\ln(\beta) = \ln(0.5) = -0.69$ . 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  $c$ , then 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 “non-parametric”. 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 “non-parametric” class makes no such distributional assumptions. Because of the confusions surrounding the parametric/non-parametric distinction, in this paper parametric approaches are termed model based, while “non-parametric” 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 “non-parametric” in the signal detection context. The term non-parametric is generally applied to statistical procedures if one or more of the following hold: (a) the variables are non-metric, either ordinal or nominal; (b) the distribution of the variables in the population is unknown. If only (b) holds then the procedures may be more accurately termed ‘distribution free’. Ordinal statistical procedures, i.e. procedures based on ranks, are also often termed “non-parametric”. 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 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 “non-parametric” sensitivity and bias measures make no such distributional assumptions, and so are distribution free. This includes both classic measures, such as 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

'model based' versus 'distribution free' is preferred to the contrast 'parametric' versus 'non-parametric'. 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 to 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, TSD*

The simplest version of TSD (Macmillan & Creelman, 1991; Swets, 1986) has a sensitivity measure,  $d'$  and a bias measure  $\beta_T$  (subscript  $T$  for TSD) that may be calculated from the proportion of hits,  $h$ , and proportion of false alarms,  $f$ . The measure  $d'$  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).

$$d' = z(h) - z(f) \quad 1,$$

where  $z(p)$  is the inverse normal probability corresponding to cumulative probability,  $p$  (Macmillan & Creelman, 1991, equation 2..10).

The likelihood ratio bias parameter,  $\beta_T$ , 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.  $\ln(\beta_T)$ , (Macmillan & Creelman, 1991, equation 2.10) is the TSD measure that is symmetric about a neutral value of zero:

$$\ln(\beta_T) = .5[z(h)^2 - z(f)^2] \quad 2.$$



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

$$\beta_{\text{optimal}} = [\pi_a/\pi_b][\text{payoff}(A|a) - \text{payoff}(A|b)]/[\text{payoff}(B|b) - \text{payoff}(B|a)] \quad 3,$$

where  $\pi(a)$ ,  $\pi(b)$  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  $z(\text{correct|stimulus } x)$  will be a linear function of  $z(\text{error|stimulus } x)$ ,

$$\begin{aligned} z(h) &= +d'_{\text{ROC}} + (1/s_T)z(f) && \text{for Response A} \\ \text{and } z(m) &= -d'_{\text{ROC}} + (1/s_T)z(cr) && \text{for Response B} \end{aligned} \quad 4,$$

where  $d'_{\text{ROC}}$  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”,  $p(\text{response B|stimulus } b)$ ; and  $m$  is probability of a “miss”,  $p(\text{response B|stimulus } a)$ . Equations 4 represent the TSD normal transformed ROC functions. Estimates of the measures  $s_T$  and  $d'_{\text{ROC}}$  can be obtained directly from the slopes and intercepts of equations 4.

Obviously, given  $d'_{\text{ROC}}$  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  $\beta_T$  is optimal for any experimenter determined bias condition. If the ROC curve arises from separate experiments for each point, then the ROC obtained from the Response B form of equation 4 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):

$$d'_e = 2d'_{\text{ROC}}[s_T/(1+s_T)] \quad 5.$$

Equations 4 describe the TSD ROC in normal – normal co-ordinates and enable graphical and statistical evaluation of the TSD model. One may obtain estimates of  $s_T$  and  $d'_{\text{ROC}}$  (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 Equation 4. One may then test whether  $s_T = \text{unity}$ , 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,  $\sigma_B$ ,  $\sigma_A$  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  $\text{var}(x-y)$  is the sum of  $\text{var}(x) + \text{var}(y)$ . 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 4 include the *implicit* assumption that all criterion variances are equal. If variances of the extreme confidence ratings are higher than those for the less extreme confidence ratings then there will be systematic deviations from the linearity predictions of equations 4. If all criterion variances are equal then 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 very similar form to 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  $\ln(\beta_L)$ , subscript  $L$  for Luce, where for any probability,  $p$ , the logit of  $p$ ,  $\text{lgt}(p)$  is given by:

$$\text{lgt}(p) = \ln[p/(1-p)] \quad 6.$$

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

$$\ln(\eta) = 0.5[\text{lgt}(h) - \text{lgt}(f)] = 0.5[\ln(h) - \ln(f) + \ln(1-f) - \ln(1-h)] \quad 7.$$

The bias measure,  $\beta_L$ , is given by :

$$\beta_L = h(1-h)/[f(1-f)]$$

$$\ln(\beta_L) = \ln(h) + \ln(1-h) - \ln(f) - \ln(1-f) \quad 8.$$

$\ln(\beta_L)$  in choice theory plays the same role as  $\ln(\beta_T)$  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, ASEs, of the measures are simple to calculate. If some measure X is given by the ratio or product of two independent probabilities, then the asymptotic standard error, ASE, of  $\ln(X)$  is given, by:

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

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:

$$ASE(2\ln(\eta)) = ASE(\ln(\beta_L)) = \sqrt{[1/n_{Aa} + 1/n_{Ab} + 1/n_{Ba} + 1/n_{Bb}]} \quad 9,$$

where  $n_{Jj}$  is the number of Responses "J" to stimulus  $j$ ;  $j = a, b$ ,  $J = A, B$

The equations for the choice ROC, equivalent to equation 4 for TSD are:

$$\lg t(h) = +2\lg t(\eta)_{ROC} + (1/s_L)\lg t(f) \quad \text{for Response A}$$

$$\lg t(m) = -\lg t(\eta)_{ROC} + (1/s_L)\lg t(cr) \quad \text{for Response B} \quad 10,$$

where  $s_L$  = ratio of variances for  $a$  and  $b$  choice theory [stimulus – criterion] representations.

Equations 10 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  $\lg t(\eta)_{ROC}$ ; and then obtain an estimate of the choice parameter at the cut-point as  $\lg t(\eta)_e$ , where

$$\lg t(\eta)_e = \lg t(\eta)_{ROC} [s_L/(1+s_L)] \quad 11.$$

It is often claimed that the logistic and normal distributions are so similar that one cannot distinguish equations 4 and 10. 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 only drops 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 (Kornbrot, 1978, 1980, 1984) found variance ratios much closer to 1 for the normal model than 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 then an *estimate* of  $A$ ,  $A'$  is given by (Macmillan & Creelman, 1991, equation 4.8):

$$A' = 1 - 1/4[f/h + (1-h)/(1-f)] \quad 12.$$

This is actually the average of the minimum and maximum possible area, 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 distribution free (Craig, 1979; Grier, 1971; Hodos, 1970; Macmillan & Creelman, 1996; Macmillan & Creelman, 1991). For this reason they will not be discussed further. The area bias measure proposed here,  $\beta_K$ , 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  $K'_A, K'_B$  as the average of the minimum and maximum possible areas between the ROC curve and the major diagonal to the left and right of the minor diagonals. In the following equations, empirical areas are indicated by unprimed symbols,  $A, K_A, K_B$ ; while *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,  $A', K'_A, K'_B, \beta'_K$ . Obviously,  $A = K_A + K_B + 0.5$  and  $A' = K'_A + K'_B + 0.5$ . Simple geometry (see Appendix) gives the following equations for  $K'_A$  and  $K'_B$ :

---

Insert Figure 2 about here

---

$$\begin{aligned}
 K'_A &= \frac{1}{4}(h-f) \left( h+f + \frac{f}{(1-f)} \right) \\
 K'_B &= \frac{1}{4}(h-f) \left( 2 - (h+f) + \frac{(1-h)}{h} \right)
 \end{aligned}
 \tag{13}$$

An approximate area based *point* bias measure is then given by:

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

$\ln(\beta'_K)$  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  $-C_{MAX}$  through 0 to  $+C_{MAX}$ . Neutral bias gives  $\ln(\beta'_K)$  equals zero, and equivalent biases towards responses A and B 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 a 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,  $A_{ROC} = 0.5 + K_A + K_B$  and the ROC area bias measure,  $\beta_K = K_A/K_B$ . If participants can give confidence ratings between 1 and  $C_{MAX}$  then the rating function has  $2C_{max}$  points indexed by the criterion  $k$ , ranging from 1 to  $2C_{max}$ . Equations for  $K_A$  and  $K_B$  are given below. Their derivation from simple geometry is given in the Appendix.

$$\begin{aligned}
 K_A &= \frac{1}{4} \sum_{k=1}^{k=C_{max}} \left[ (h_{k+1} - f_k)^2 - (h_k - f_{k+1})^2 \right] \\
 K_B &= \frac{1}{4} \sum_{k=C_{max}+1}^{k=2C_{max}} \left[ (h_{k+1} - f_k)^2 - (h_k - f_{k+1})^2 \right]
 \end{aligned}
 \tag{15}$$

Equations 15 can then be used to give a more accurate area bias measure at the ‘‘A’’, ‘‘B’’ cut point, denoted  $\beta_K$ , analogous to the measure in equation 14:

$$\beta_K = K_A/K_B
 \tag{16}$$

The  $\beta_K$  measure behavior at the "A", "B" cut point, but use 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 4 possible responses) the underestimation would be substantial. Hence  $A_{ROC}$  and  $\beta_K$  should only be used 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 & 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) considers 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 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) then defines the sensitivity measure,  $S'$  as the sum of  $U_R(k)$  over all criteria:

$$S' = \sum_{k=1}^{2C_{\max}} U_R(k) \quad 17.$$

When the number of criteria is small, as for example with a Likert scale where people rate responses, as 'uncertain', 'moderately confident', 'very confident', most (but not necessarily all) participants use all responses. However, when the confidence is on a rating scale from 1-100 or on a continuous slider, then people rarely use all possible responses and the measure  $S'$  is confounded by the *number* of different responses a person chooses to use. This measure is provided here as it is presented as an alternative to  $d'$  by Balakrishnan and his co-workers, (Balakrishnan, 1998a, 1998b, 1999). Balakrishnan himself no longer promotes  $S'$  (personal communication January, 2004).

Instead, he and his co-workers suggest a measure  $\gamma_0$  that estimates unbiased performance (Balakrishnan, MacDonald, & Kohen, 2003). Whether  $\gamma_0$  remains constant across conditions with different a priori stimulus frequencies and/or different payoffs does not appear to have been tested.

Hence  $\gamma_0$  is not considered here.

Balakrishnan and his co-workers, (Balakrishnan, 1998a, 1998b, 1999) also note 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. They suggest that a useful bias measure should assess both the shift of the whole ROC function *and* be independent of the distribution of the stimulus representation. With these goals in mind, they define sub-optimal, in their terms biased, responses as responses where  $p(\text{correct} | R=k) < 0.5$ . For such a criterion,  $k$ , an estimate of bias  $\omega_k$  is equal to the total number of responses at that value of  $k$  divided by the total number of trials (Balakrishnan, 1998b, footnote to Table 5).

Then:

$$\begin{aligned}
 \omega_{kA} &= 0 && \text{if } p(A|a) > p(A|b) \\
 \omega_{kA} &= \left[ \frac{n(A|a) + n(A|b)}{N} \right] && \text{if } p(A|a) < p(A|b) \\
 \Omega_A &= - \sum_{k=1}^{k=C_{\max}} \omega_{kA} \\
 \omega_{kB} &= 0 && \text{if } p(B|b) > p(B|a) \\
 \omega_{kB} &= \left[ \frac{n(B|b) + n(B|a)}{N} \right] && \text{if } p(B|b) < p(B|a) \\
 \Omega_B &= + \sum_{k=C_{\max}+1}^{k=2C_{\max}} \omega_{kB}
 \end{aligned}
 \tag{18}$$

where total number of trials is  $N$ ,  $\Omega_A$  is a measures of bias on occasions when response “A” is given, and is negative; while  $\Omega_B$  is a measures of bias on occasions when response “B” is given, and is positive. The total bias of the complete ROC function is then defined as  $\Omega_T = \Omega_A + \Omega_B$ . Positive values of  $\Omega_T$  indicate bias towards “ B” and vice versa. Balakrishnan and his colleagues’ approach to sub-optimality and bias is thus different to 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  $\omega_k$  as a measure the magnitude of that bias. The  $\omega_k$  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  $\Omega_T$  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  $\Omega_T$  will behave differently from other bias measures.

*The ln(odds ratio) as a Distribution Free Bias Measures*

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 their sample estimates are generated by a binomial distribution. The ln(odds ratio) is identical with the choice bias measure  $\ln(\beta_L)$ .

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, then 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 then the function should be displaced. The general prediction is:

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

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

The constant should be zero for neutral conditions and negative for bias towards response B. The constants  $G_{\text{measure}}$  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:  $\beta_T$  from equation 2; for  $\beta_L$  from equation 8; and for  $\beta'_K$  from equation 14. (One might also use  $\beta_K$  from equation 16, but it is so similar to  $\beta'_K$  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 stimuli, see equation 3. Equations 19 are important because they provide a way to evaluate the bias of the whole *ROC* function that depends on cumulative probabilities, unlike  $\Omega$  that 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  $-1$ , then there is symmetry about the point  $G_{\text{measure}}$ , and one can state that participant is consistent in 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 responses “A” and “B”.

### *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, ( $k = -C_{MAX}, -C_{MAX} + 1 \dots 0, C_{MAX}-1, C_{MAX}$ ). 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:

$$S_{rel} = (\text{sensitivity at criterion } k) / (\text{sensitivity at the cut-point}) \quad 20.$$

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 any way, even if they are statistically significant. The second criterion is the frequency 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 percentage overestimation and underestimation of sensitivity relative to the cut point sensitivity, where

$$\text{Over estimate } \% = 100(\text{sensitivity at criterion } k) / (\text{sensitivity at the cut-point}) - 100$$

$$\text{Under estimate } \% = 100 - 100(\text{sensitivity at criterion } k) / (\text{sensitivity at the cut-point}) \quad 21,$$

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 are compared with their *ROC* equivalents for TSD, choice and area formulations.

Finally, the invariance of sensitivity measures across different bias conditions is 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 equations 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 only makes sense 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, i.e. is response “B” made more often if a correct response to stimulus *b* is more highly rewarded, or if stimulus *b* has 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 4 (TSD) and equations 10 (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 (1999) Data

The data analyzed here is described by Balakrishnan (1999), and is available on-line (Balakrishnan & MacDonald, 2003a). It comprises individual data on four participants performing a difficult line length discrimination task, with confidence ratings. All four participants performed in both a neutral and a biased condition. Experiment 2 was a frequency manipulation experiment. There were two 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 where the frequency of stimulus *b* was 3 times the frequency of stimulus *a*. Experiment 3 was a payoff manipulation experiment. There were two 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 where the rewards for the correct response and penalty for the wrong response, to stimulus *b* were 3 times the rewards and penalties for equivalent response to stimulus *a*. Optimal bias towards “A” is  $1/3$ , in both the frequency and payoff biased conditions. The neutral condition came first in both experiments. Allowed confidence ratings were 1-100, derived from a slider scale. All analyses are 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 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 performance lower than 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  $D_{pay}$  are all appropriately biased towards response “B” in the biased condition. However  $C_{pay}$  was just as biased towards response “B” in the neutral condition. Thus, three out of four participants do show voluntary control of bias. Two out of these three,  $A_{freq}$  and  $D_{pay}$ , appear to change 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*

---

Insert Figures 3 and 4 and Table 1 about here

---

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 3 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 the other measures, for all participants except  $D_{pay}$ . For  $D_{pay}$  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 models. Furthermore the % deviations from cut-point values were much larger for underestimates than overestimates. The final two columns of Table 1 show these % deviations separately for underestimates and overestimates, according to equation 21. Here the  $A'$  measure is numerically superior for all participants in all conditions. This superiority of  $A'$  was a surprise, as 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 very close to the identity relation with adjusted r-squared greater than 99% and slope unity and intercept zero. For TSD: slope = 1.05 (.98, 1.12), intercept = -.07 (-.19, .05). For choice: slope = 1.02 (.94, 1.10), intercept = -.07 (-.16, .05). The area measure *ROC* and *point* measures were slightly less similar. Adjusted r-squared equals .971, slope = 1.25 (1.05, 1.45), intercept = -.24 (-.40, .057). So overall this data shows almost identical performance for *ROC* and *point* sensitivity measures, and hence no advantage for the more complex confidence rating procedure.

---

Insert Figure 5 about here

---

### *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 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 three people performed better in the biased condition and one performed (very slightly) better in the neutral condition. Balakrishnan's  $S'$  shows a different pattern. It is larger, sometimes much larger, for the neutral condition for all participants. The reason is simple. 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).

---

Insert Table 2 about here

---

### *Performance of the Bias Measures*

The behavior of  $\ln(\beta_K)$  and  $\ln(\beta'_K)$  are almost identical. For all four 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 not significantly different from zero. Hence only the more easily calculated  $\ln(\beta'_K)$  is used in most of what follows.

*Consistency of point bias measures as a function of criterion across responses “A” and “B”*

---

Insert Figure 6 and Table 3 about here

---

Figures 6 and 7 examine the consistency of usage of the confidence level for the bias measures  $\ln(\beta_T)$ ,  $\ln(\beta_L)$ ,  $\ln(\beta'_K)$ , by testing equations 19. Table 3 shows intercepts =  $G_{\text{measure}}$ , 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 for Cpay in the biased condition for all 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 analyses is due to the very restricted range for response “A”.

*Comparison of point and ROC bias measures*

Figure 8 shows *ROC* bias  $G_{\text{measures}}$  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. Firstly, for all participants and all conditions, *ROC* measures are considerably (at least 1.6 times) larger than the equivalent *point* measure. Secondly, the behavior of Cpay in the biased condition is substantially different from that of the other three participants, as is also evident in Figure 7. If one excludes Cpay in the biased condition, then regressions in Figure 7 all have intercepts not significantly different from zero at the 95% confidence level. The regression slopes are: *area* = 1.64 (1.37, 1.92); TSD = 2.18 (1.84, 2.51); choice = 2.13 (1.82, 2.45).

---

Insert Figure 7 about here

---

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 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 other participants) together with a rather small *point* criterion shift (less than other participants).

#### *Comparison of Neutral and Biased Conditions: Voluntary Control of Bias Measures*

Table 4 shows *point* measures of bias,  $\ln(\beta_T)$ ,  $\ln(\beta_L)$ ,  $\ln(\beta'_K)$ ,  $\ln(\beta_K)$  at the cut-point between responses “A” and “B”, together with *ROC*  $G_{\text{measures}}$  from equations 19 and Balakrishnan’s  $\Omega$  for each participant in a neutral and a biased condition. All measures except  $\Omega$  show *different* behavior in the biased and neutral conditions, and hence voluntary control. For the *point* measures, statistical tests of these differences are available for choice model using the ASEs in equation 9; and all differences are statistically significant at the 95% confidence level. For the  $G_{\text{measures}}$  standard errors are available from the regressions in equations 19. Again neutral and biased conditions show significant differences for all participants.

---

Insert Table 4 about here

---

#### *Optimality of the Bias Measures*

Table 4 shows little support for any strong version of optimality. The only person with a bias parameter not significantly different from 0 in the neutral conditions is *Bfreq*. The only measure not significantly different from 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 measures except  $\Omega$ . Table 4 show 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 towards “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” to “B” cut-point (where a test is possible because the ASE is available). As noted above, *Bfreq* moves in the normatively wrong direction on all measures. For *point* measures, the results for *Cpay* are equivocal, being small towards “A” for TSD and choice, but small towards

“B” for the *area* measure.

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

#### *Fit of TSD and Luce's Choice Model*

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

The fit of TSD and choice model is first evaluated by testing the linearity of their respective ROC functions separately for response “A” and response “B”, for all participants in both conditions. There is little to choose between TSD and choice. Both models had 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 non-linear 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 equations 4 for all TSD functions except *Dpay* in the neutral condition (equation 4), and for all choice functions except *Dpay* in both neutral and biased conditions (equation 10). 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.

---

Insert Figures 9 and 10 and Table 5 about here

---

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 4 for TSD and 10 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 8 estimates of  $s_T$  are significantly less than unity, mean  $s_T = 0.79$ , (99% confidence limits .72, .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  $s_L$  is .99, (99% confidence limits .87, 1.10). Thus the equal variance version of choice model is viable for some 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*

Based on the analyses presented here, as well as 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,  $\ln(\beta'_{\kappa})$ , for simple two choice experiments, even without confidence ratings. Given rating data, one can compare models based on different distributions.

### *Sensitivity*

#### *Consistency*

All measures of *point* sensitivity showed an effect of confidence rating within conditions. Figures 3 and 4 show there is a substantial middle range of confidence ratings where sensitivity remains constant, and very similar to the value at the A to B cut-point. Sensitivity values estimated from very extreme confidence ratings are lower than those at the cut-point. Surprisingly, as documented by measures in Table 1, the  $A'$  measure shows the *least* variation over different confidence ratings. This *might* be an artifact of the fact that the range of  $A'$  is limited from 0.5 to 1.0, while  $\ln(\eta)$  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 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  $S'$ . Although Balakrishnan argues that such unequal use does not make much difference (Balakrishnan & MacDonald, 2002), he no longer promotes use of  $S'$  (personal communication). 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 neutral conditions for the *area* and model based measures. These favored the biased condition (performed second for all participants) for three out of the four participants.

#### *Bias*

##### *Consistency*

Figures 6 and 7, based on equations 19, provide a new way of measuring consistency of criterion 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 7 out of 8 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  $G_{\text{measure}}$ . Armed with this measure, one can compare *point* and *ROC* bias measures. Values of  $\beta'_K$  from equation 14 and  $\beta_K$  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  $G_{\text{measure}}$  for the entire *ROC* function with equivalent *point* measures gives a different picture. As shown in Table 4, the *ROC* measures are always higher. Even the

neutral conditions show some degree of bias using  $G_{\text{measure}}$ . It thus appears that the full ROC is more sensitive to deviations from neutrality than 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  $C_{\text{pay}}$  is only detectable in Figure 7, and by the high values of  $G_{\text{measure}}$  in Table 3.

By contrast, Balakrishnan's  $\Omega$ , also a full ROC measure, shows minimal bias for all participants in all 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 measures, except  $\Omega$  show voluntary control of bias, in that values of at least some bias measures are different in the neutral and biased conditions. In terms of *point* measures,  $A_{\text{freq}}$ ,  $C_{\text{pay}}$  (minimally) and  $D_{\text{pay}}$  all change their bias in the normatively correct direction, while  $B_{\text{freq}}$  changes in the opposite direction. This pattern is observed for all 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  $G_{\text{measure}}$ 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 choose this optimal bias. In terms of the raw ratio of proportion of "A" response relative to proportion of "B" responses, people with lower sensitivities need to be *more* biased. This is not happening, or not happening sufficiently, so the least sensitive participant,  $A_{\text{freq}}$  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 sub-optimal without such training.

In real world applications different situations may have both different a priori probabilities and different payoffs. For example, malignant cells in screening conditions are less frequent 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, etc. Similar arguments apply to the probability and degree of threat of different kinds of military weapons, or

computer viruses, or of risk of different 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.

#### *Model Evaluation and ROC Functions*

Balakrishnan (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2003b) suggests 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 fit of both TSD and choice does 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. Sub-optimality 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, 0.79, is substantially less than unity, suggesting that the stimulus *a* distribution has higher variance than the stimulus *b* distribution. Choice model is to be preferred to 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 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  $\Omega$  emerge from these analyses.

Nevertheless, model based approaches are clearly essential for deeper understanding of underlying processes. For example, TSD or choice measure 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.

- The signal detection approach is successful and useful, not flawed as suggested by Balakrishnan and his colleagues
- *Area* measures are the best distribution free measures. The new *area* bias measure,  $\ln(\beta'_{\kappa})$ , complements the well established sensitivity measure  $A'$
- *Area* measures are at least as good as TSD or choice for practical purposes.
- *Point* measures of sensitivity and of bias at the cut-point are just as good as *ROC* measures, and much simpler to obtain.
- The new techniques for assessing the bias of complete ROCs are important and give more, and different, information to that from bias at the cut-point alone.
- There are small but significant departures from the predictions of choice and TSD.
- The simpler equal variance version of choice model is acceptable, while 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 TSD.

### References

- Agresti, A. (1996). *Introduction to categorical data analyses*. Chichester: Wiley.
- Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, 40(4), 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 and Performance*, 25(5), 1189-1206.
- Balakrishnan, J. D., & MacDonald, J. A. (2002). Decision criteria do not shift: Reply to Treisman. *Psychonomic Bulletin and Review*, 9(4), 858-865.
- Balakrishnan, J. D., & MacDonald, J. A. (2003a). Alternatives to Signal Detection Theory. Retrieved 17 Aug 2003, 2003, from <http://www.psych.purdue.edu/~beowulf/dsdt/dsdt.html>
- Balakrishnan, J. D., & MacDonald, J. A. (2003b). Decision criteria do not shift: Reply to Treisman. *Perception & Psychophysics*.
- Balakrishnan, J. D., MacDonald, J. A., & Kohen, H. S. (2003). Is the area measure a historical anomaly? *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale*, 57(3), 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. *Perception-and-Psychophysics*, 17, 167-178.
- Dusoir, A. E. (1983). Isobias curves in some detection tasks. *Perception and 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* (Vol. 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 and Statistical Psychology*, 41(1), 119-144.

- Hodos, W. (1970). A non parametric 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-and-Psychophysics*, *63*, 1091-1100.
- Kornbrot, D. E. (1978). Theoretical and empirical comparison of choice models. *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*, 441-558.
- 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. (1986). *Response times*. Oxford: 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*. New York, NY, US: John Wiley & Sons, In.
- Macmillan, N. A., & Creelman, C., D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic-Bulletin-and-Review*, *3*, 164-170.
- Macmillan, N. A., & Creelman, D. C. (1991). *Detection theory: a user's guide*. Cambridge: Cambridge University Press.

- McCarthy, D., & Davison, M. (1981). Towards a behavioral theory of bias in signal detection. *Perception-and-Psychophysics*, *29*, 371-382.
- McCarthy, D., & Davison, M. (1984). Isobias and alloibias functions in animal psychophysics. *Journal-of-Experimental-Psychology:-Animal-Behavior-Processes*, *10*, 390-409.
- Pollack, I., & Norman, D. A. (1964). A nonparametric 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 and 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.
- Vickers, D., Caudrey, D., & Willson, R. J. (1971). Discriminating between the occurrence of two alternative events. *Acta Psychologica*, *35*, 151-172.



Table 1

*Performance of Point Sensitivity Measures as a Function of Confidence Criteria*

Participant	Condition	# Criteria	Sensitivity	# Outside	Max %	Max %
				15%	Overestimate	Underestimate
Afreq	Neutral	97	TSD: $d'$	33	2.5	82.9
			Choice: $\ln(\eta)$	29	2.4	74.6
			Area: $A'$	3	0.6	31.7
	Biased	60	TSD: $d'$	22	0.3	91.7
			Choice: $\ln(\eta)$	20	0.4	86.1
			Area: $A'$	8	0.1	23.6
Bfreq	Neutral	101	TSD: $d'$	20	10.4	38.8
			Choice: $\ln(\eta)$	16	46.1	29.1
			Area: $A'$	-	0.5	11.8
	Biased	79	TSD: $d'$	25	3.4	42
			Choice: $\ln(\eta)$	11	4.2	25.9
			Area: $A'$	-	0.1	13.6
Cpay	Neutral	97	TSD: $d'$	29	4.9	36
			Choice: $\ln(\eta)$	21	50.5	46.9
			Area: $A'$	-	0.1	12.4
	Biased	85	TSD: $d'$	18	0.5	48.2
			Choice: $\ln(\eta)$	12	1	24.9
			Area: $A'$	3	0.1	16.1
Dpay	Neutral	119	TSD: $d'$	16	0.2	30.6
			Choice: $\ln(\eta)$	-	11.2	4.6
			Area: $A'$	4	< .05	15.8
	Biased	38	TSD: $d'$	5	0.4	32.6
			Choice: $\ln(\eta)$	1	16	13.1
			Area: $A'$	2	0.1	15.9

Table 2

*Point and ROC Sensitivity Measures in Neutral and Biased Conditions*

Analysis	Participant	Point		ROC	
		Neutral	Bias	Neutral	Bias
		$d'$		$d'e$	
TSD	Afreq	.67	.63	.65	.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	.54	.52	.51	.51
	Bfreq	1.01	1.18	.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
		$\Omega$			
Balakrishnan	Afreq			20.1	8.8
	Bfreq			37.6	32.4
	Cpay			37.8	31.8
	Dpay			70.4	20.9

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.*

Participant	Analysis	$G_{measure}$		Slope		Adjusted $r^2$	
		Neutral	Bias	Neutral	Bias	Neutral	Bias
Afreq	Area	.58	-1.36	-.48	-.64	.991	.997
	Choice	.22	-.50	-.50	-.99	.895	.275
	TSD	.17	-.40	-.50	-.98	.895	.311
Bfreq	Area	.02	.36	-.68	-.98	.978	.994
	Choice	-.14	.42	-1.29	-.87	.983	.987
	TSD	-.11	.34	-1.28	-.88	.985	.988
Cpay	Area	-1.04	-10.67	-.69	-19.15	.997	.776
	Choice	-1.19	-9.16	-.84	-12.12	.994	.628
	TSD	-.92	-7.50	-.79	-12.31	.992	.641
Dpay	Area	-.07	-.55	-.64	-.51	.981	.929
	Choice	-.31	-3.05	-.96	-1.35	.989	.886
	TSD	-.25	-2.45	-.94	-1.29	.989	.893

Table 4

*Point and ROC Bias Measures in Neutral and Biased Conditions*

Analysis	Participant	<i>Point</i>			<i>ROC</i>			Full ROC from		
		at "A","B" cut			at "A","B" cut			Equation 19		
		Neutral	Bias	B-N	Neutral	Bias	B-N	Neutral	Bias	B-N
		$\ln(\beta_T)$						$G_{TSD}$		
TSD	Afreq	.12	-.23	-.35				.17	-.40	-.57
	Bfreq	-.01	.17	.18				-.11	.34	.45
	Cpay	-.53	-.62	-.09				-.92	-7.50	-6.58
	Dpay	-.12	-1.08	-.96				-.25	-2.45	-2.20
		$\ln(\beta_L)$						$G_{choice}$		
Choice	Afreq	.15	-.29	-.44				.22	-.50	-.72
	Bfreq	-.01	.22	.23				-.14	.42	.56
	Cpay	-.66	-.77	-.11				-1.19	-9.16	-7.97
	Dpay	-.14	-1.31	-1.17				-.31	-3.05	-2.74
		$\ln(\beta'_{A'})$			$\ln(\beta_A)$			$G_{area}$		
Area	Afreq	.38	-.80	-1.18	.24	-.72	-.97	.58	-1.36	-1.93
	Bfreq	-.02	.14	.16	-.12	.10	.22	.02	.36	.34
	Cpay	-.62	-.56	.06	-.57	-.64	-.07	-1.04	-10.67	-9.63
	Dpay	-.07	-.47	-.40	-.08	-.54	-.47	-.07	-.55	-.48
								$\Omega$		
Balakrishnan	Afreq							.06	.00	-.06
	Bfreq							.01	.02	.01
	Cpay							.04	.00	-.04
	Dpay							.00	.00	.00

Note a. B-N is value in biased condition – value in neutral condition. Negative values indicate a shift towards response B.

Table 5

*Tests of equal variance versions of TSD and choice model*

Participant	Neutral		Bias	
	slope <sup>a</sup>	se(slope) <sup>b</sup>	slope <sup>a</sup>	se(slope) <sup>b</sup>
<i>s<sub>T</sub> from TSD</i>				
Afreq	.79	.007	.81	.009
Bfreq	.87	.018	.78	.014
Cpay	.89	.012	.64	.013
Dpay	.76	.013	.79	.059
Mean	.83		.75	
LCL <sup>c</sup>	.77		.69	
UCL <sup>c</sup>	.89		.81	
<i>s<sub>L</sub> from choice</i>				
Afreq	.85	.008	.87	.012
Bfreq	<b>1.08</b>	.026	<b>.99</b>	.019
Cpay	<b>1.17</b>	.021	.81	.013
Dpay	<b>1.03</b>	.012	<b>1.09</b>	.072
Mean	<b>1.03</b>		<b>.94</b>	
LCL <sup>c</sup>	.95		.86	
UCL <sup>c</sup>	1.11		1.02	

Note a. Slopes *not* significantly different from 1 at the 99% confidence level are in bold.

Note b. Standard errors of slopes are obtained by averaging standard errors obtained from the stimulus *a* and stimulus *b* regressions.

Note c. Lower confidence levels (LCL) and upper confidence levels (UCL) are based on standard errors that are averages of the 8 values of (standard error)<sup>2</sup> for the 4 participants in 2 conditions each.

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 co-ordinates  $f, h$ , 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 (triangle OXC), plus *half* the triangle OVC that would need to be added to OXC to obtain the maximum area. This is the same pragmatic approach used by Pollack and Norman (Pollack & Norman, 1964), see also (Macmillan & Creelman, 1996). So,

$$K'_A = \text{area } \Delta OXC + .5 \text{ area } \Delta OVC.$$

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 (triangle  $\Delta IXC$ ), plus *half* the triangle  $\Delta IVC$  that would need to be added to IXC to obtain the maximum area, thus:

$$K'_B = \text{area } \Delta IXC + .5 \text{ area } \Delta IVC.$$

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

$$x = (h + f)/\sqrt{2}, \quad y = (h - f)/\sqrt{2}$$

The value of  $u$  may be obtained from the similar triangles  $\Delta OV'C, \Delta OMU$  by noting that:

$$(1 - u)/f = 1/h, \quad \text{therefore } u = 1 - fh.$$

Similarly, the value of  $v$  may be obtained from the similar triangles  $\Delta IU'C, \Delta IMV$  by noting that:

$$(1 - v)/(1 - h) = 1/(1 - f), \quad \text{therefore } v = 1 - (1 - h)(1 - f).$$

Using these values for the distances  $x, y, u, v$  give the following equations for  $K'_A, K'_B$ :

$$\begin{aligned} K'_A &= \text{area } \Delta XOC + .5 \text{ area } \Delta VOC = \frac{xy}{2} + \frac{1}{2} \frac{vf}{2} \\ &= \frac{h^2 - f^2}{4} + \frac{f(h - f)}{4(1 - f)} \\ &= \frac{(h - f)}{4} \left( h + f + \frac{f}{1 - f} \right) \end{aligned} \quad 22.$$

$$\begin{aligned}
K'_B &= \text{area}\Delta ICX + .5\text{area}\Delta ICU = \frac{(\sqrt{2-x})y}{2} + \frac{1}{2} \frac{u(1-h)}{2} \\
&= \frac{2(h-f)}{4} - \frac{h^2 - f^2}{4} + \frac{(h-f)(1-h)}{4h} \\
&= \frac{(h-f)}{4} \left( 2 - (h+f) + \frac{(1-h)}{h} \right)
\end{aligned} \tag{23}$$

*Areas Under the Empirical ROC Curve,  $K_A$ ,  $K_B$*

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  $2 C_{MAX}$  criteria,  $k$  from 1 to  $2 C_{MAX}$ . 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, the major diagonal and to the left of the cut line  $CX_c$ .  $K_B$ , striped on the figure, is the area bounded by the empirical ROC function, the major diagonal and to the right of the cut line  $CX_c$ . The  $k$ th 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  $X_k P_k P_{k+1} X_{k+1}$ . The distance  $OX_k$  along the major diagonal is denoted  $x_k$  and the distance  $X_k P_k$  parallel to the minor diagonal is denoted  $y_k$ . Then,

$$x_k = (h_k + f_k)/\sqrt{2}; \quad y_k = (h_k - f_k)/\sqrt{2} \tag{24}$$

The polygon  $X_k P_k P_{k+1} X_{k+1}$  is composed of the rectangle  $X_k P_k T X_{k+1}$  plus the triangle  $P_k P_{k+1} T$ , and so has an area given by:

$$\text{area } X_k P_k P_{k+1} X_{k+1} = y_k(x_{k+1} - x_k) + [(y_{k+1} - y_k)(x_{k+1} - x_k)]/2 = [(h_{k+1} - f_k)^2 - (h_k - f_{k+1})^2]/2.$$

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

$$\begin{aligned}
K_A &= \frac{1}{4} \sum_{k=1}^{k=c} \left[ (h_{k+1} - f_k)^2 - (h_k - f_{k+1})^2 \right] \\
K_B &= \frac{1}{4} \sum_{k=c+1}^{k=2CR+1} \left[ (h_{k+1} - f_k)^2 - (h_k - f_{k+1})^2 \right]
\end{aligned} \tag{25}$$

## Figure Captions

*Figure 1.* ROC function  $p(\text{hit}), h$ , as a function of  $p(\text{false alarm}), f$ , for all participants. Top panel neutral conditions, bottom panel biased conditions. Filled triangles from response “A”, open squares from response “B”. Top left panel shows key areas shaded, as an example, with  $\Delta\text{OCX} = K_A$ , dotted (corresponding to response “A”);  $\Delta\text{ICX} = K_B$ , striped (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.

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

*Figure 3.* 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  $A_{\text{freq}}$ , right panels  $B_{\text{freq}}$ . Top panels relative  $d'$ , middle panels relative  $\ln(\eta)$ , bottom panels relative  $A'$ .

*Figure 4.* 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  $C_{\text{pay}}$ , right panels  $D_{\text{pay}}$ . Top panels relative  $d'$ , middle panels relative  $\ln(\eta)$ , bottom panels relative  $A'$ .

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

*Figure 6.* Response “B” bias measure as function of response “A” bias measure for frequency manipulation participants.. Top panels Area measure,  $\ln(\beta'_K)$ ; middle panel TSDmeasure  $\ln(\beta_T)$ , bottom panel choice measure  $\ln(\beta_L)$ .



*Figure 7.* Response "B" bias measure as function of response "A" bias measure for payoff manipulation participants.. Top panels *area* measure,  $\ln(\beta'_K)$ ; middle panel TSDmeasure  $\ln(\beta_T)$ , bottom panel choice measure  $\ln(\beta_L)$ .

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

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

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





















