Notes and Comment

Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999)

MICHEL TREISMAN

University of Oxford, Oxford, England

For nearly 50 years, signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991) has been of central importance in the development of psychophysics and other areas of psychology. The theory has recently been challenged by Balakrishnan (1998b), who argues that, within SDT, an alternative index is “better justified” than d' and who claims to show (1998a, 1999) that SDT is fundamentally flawed and should be rejected. His evidence is based on new nonparametric measures that he has introduced and applied to experimental data. He believes his results show that basic assumptions of SDT are not supported—in particular, that payoff and probability manipulations do not affect the position of the decision criterion. In view of the importance of SDT in psychology, these claims deserve careful examination. They are critically reviewed here. It appears that it is Balakrishnan’s arguments that fail, and not SDT.

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 1991; Tanner & Swets, 1954) has been of major importance in the development of psychophysics and other areas of psychology. The account it provides of the processing of sensory information and the operation of the sensory decision-making process is now built into the fabric of much of experimental psychology. In three recent papers, Balakrishnan (1998a, 1998b, 1999) has claimed, however, that an alternative index is “better justified” than d' and that SDT is fundamentally flawed and must be rejected. His argument is based on the use of new measures that he has developed and applied experimentally. Although he states that his “new measures are rooted within the general framework of statistical decision theory, and hence are also consistent with earlier empirical results, including ROC curve data” (Balakrishnan, 1998b, p. 69), his account also “does not require the technical assumptions of signal detection theory” (Balakrishnan, 1998a); he stresses that it has the advantage of being distribution free. His results present a major challenge to the assumption, central to SDT, that variation in probabilities or payoffs can cause the position of the decision criterion to shift.

However, if new procedures produce results that conflict with previous findings, two possibilities must be considered: (1) The accepted interpretation has been shown to be wrong, or (2) there are problems with the new procedures. Balakrishnan reaches the first conclusion. Here, I will examine the validity of his measures and assess the weight his conclusions deserve.

The main measures or tests are (1) S' , a measure of sensitivity that is claimed to be “better justified and statistically more reliable” (Balakrishnan, 1998b, p. 69) than d' , (2) a procedure for determining the true position of the criterion, using his measure $U_R(k)$, (3) a measure of response bias Ω_p , which is used to determine whether the decision criterion departs from the unbiased position, and (4) a procedure for detecting suboptimality and, thus, revealing that the decision criterion is not at the optimal position given by SDT. In each case, these measures are claimed to give results incompatible with SDT. There are relations between the measures, but for simplicity, they will be discussed separately.

I will commence by presenting a simple SDT model and deriving a relation between parameters from it; this will serve to provide a basis for the analysis and discussion of Balakrishnan’s new measures. I will then consider each of the measures in turn. In each case, a comparison will be made with relevant predictions derived from SDT. Finally, I will consider whether his results present any serious problems for SDT. (The notation sometimes varies between Balakrishnan’s different papers. For simplicity, a single notation is used here.)

Signal Detection Theory and d'

SDT measures have been derived for and applied to an extensive range of procedures and problems (Green & Swets, 1966; Macmillan & Creelman, 1991). Balakrishnan’s (1998a, 1998b, 1999) new methods, however, apply only when the rating procedure is used to discriminate between two stimuli. (These stimuli will be referred to here as the noise stimulus, n , and the signal stimulus, s , and the responses as *no* and *yes*.) The present discussion will be restricted to the rating procedure and will commence by presenting a simple SDT model for it. Figure 1 illustrates the model as applied to a detection or discrimination task in which two stimuli occur in random sequence and the subject must choose between two responses.

In Figure 1B, density functions $f_n(x)$ (produced by presentations of n , the noise stimulus) and $f_s(x)$ (produced by presentations of s , the signal), with means of $\mu_n = 0$ and $\mu_s = a$ and standard deviations of $\sigma_s = \sigma_n = a/2$, are plotted on the decision axis scale x (continuous lines). An alternative signal density function, $f'_s(x)$, with a mean of $\mu_s = a$ and a standard deviation of $\sigma_s = a/2^{1/2}$ (dashed

I acknowledge the support extended by the Leverhulme Trust. I also thank Ivan Smith for useful comments. Correspondence may be addressed to M. Treisman, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England (e-mail: treisman@psy.ox.ac.uk).

Figure 1. In panel B, normal density functions $f_n(x)$ and $f_s(x)$ with means $\mu_n = 0$ and $\mu_s = a$ and standard deviations $\sigma_s = \sigma_n = a/2$ describe the sensory effects produced on a central decision axis x by the noise and signal stimuli (continuous lines). An alternative signal density function $f'_s(x)$, with mean $\mu_s = a$ and $\sigma_s = a/2^{1/2}$ (dashed line) is also shown. A decision criterion $x_c = a/2$ on the decision axis has likelihood ratio $l(x_c) = 1$. In panel A, the corresponding distribution functions, $F_n(x)$ and $F_s(x)$ (continuous lines) and $F'_s(x)$ (dashed line), relate $P(\text{NO})$ to the position of the criterion x_c on the decision axis.

line), is also shown. These distributions are taken here to be normal (although SDT does not require this). They describe the sensory effects produced on a central decision axis by repeated presentations of the noise and signal stimuli. The decision criterion shown is $x_c = a/2$. Since it lies at the intersection of the density functions, its likelihood ratio is $l(x_c) = f_s(x_c)/f_n(x_c) = 1$. A criterion with a likelihood ratio of 1 may be described as unbiased; it is the appropriate value for equal probabilities of signal and noise and equal values and costs for all pay-offs. Criteria at all other positions are described as biased. Incoming sensory inputs falling below x_c evoke the *no* response; those above the criterion, *yes*. Figure 1A presents the corresponding distribution functions, $F_n(x)$ and $F_s(x)$ (continuous lines) and $F'_s(x)$ (dashed line).

These plot $P(\text{NO})$, the probability of the *no* response, as a function of the position of the criterion on the decision axis, for presentations of the respective stimuli. For the criterion $x_c = \mu_n = 0$, $P(\text{NO} | n)$, the probability of a correct rejection, is .5, and $P(\text{NO} | s)$, the probability of a miss, is .0223. Vertical lines 0c and ab, at 0 and a , are also shown. These lines intersect $F_n(x)$ and $F_s(x)$ at the 50 percent response points $(0, p_{0.5})$ and $(a, p_{0.5})$, respectively.

SDT allows us to estimate the parameters of the underlying distributions and to use these estimates to provide a measure of the sensitivity of the subject. The efficiency of detection or discrimination increases as the overlap between the density functions decreases, either because they become more widely separated on the decision axis or because their variances are reduced. Thus, an estimate of the separation between the means of the distributions might provide a first approximation to a measure of sensitivity. We may define the separation between the means as

$$\mu_{\text{dif}} = \mu_s - \mu_n. \quad (1)$$

For the density functions $f_n(x)$ and $f_s(x)$ in Figure 1, $\mu_{\text{dif}} = \mu_s - \mu_n = a$. At first sight, it might seem that μ_{dif} might be estimated by procedures that fit the functions $F_n(x)$ and $F_s(x)$ and determine the difference between their 50 percent points. However, this is not feasible, and even if it were, it would not be satisfactory, for the following reasons.

First, we cannot determine scale values on the decision axis x directly, and we do not know what an appropriate scale would be. (Would it correspond to the rate of firing of unidentified neurons in the nervous system, or to some other variable?) Thus, lacking independent experimental access to it, we cannot use the decision axis scale as an independent variable.

Second, even if we could identify an appropriate scale for the decision axis and were able to measure differences on this scale, we would not be able to use these differences as unambiguous measures of the system's sensitivity. The problem is that for a given value of μ_{dif} , sensitivity decreases if σ^2 increases, but the measure takes no account of this.

The scale problem arises because we cannot compare a distance on x to an external standard, so as to assign a measure to that distance in a repeatable way, as we might use a ruler to determine that a line is a given number of centimeters long. However, a solution to this problem is built into SDT. This defines an origin for the decision axis scale by setting $\mu_n = 0$ and a unit by setting $\sigma_n = 1$. SDT procedures then allow an estimate of the normalized difference between the means to be used as a measure of discrimination:

$$d' = \frac{\mu_{\text{dif}}}{\sigma_n} = \frac{\mu_s - \mu_n}{\sigma_n}. \quad (2)$$

This normalizes the separation between the density functions by division by the noise standard deviation, so giving a unique scale for the decision axis on which d' provides a measure of sensitivity. This measure can be used

Figure 2. Two step functions, $S_0(x)$ and $S_a(x)$ are shown, plotted against the decision axis, with means of 0 and a , as in Figure 1A. In addition a ramp function $L_0(x)$, with a mean of 0, and $G_a(x)$, a monotonic distribution function with odd symmetry and a mean of a , are shown. The area between $S_0(x)$ and $S_a(x)$ is equal to the area between $L_0(x)$ and $S_a(x)$ (since the triangles T_1 and T_2 are equal in area) and approximately equal to the area between $S_0(x)$ [or $L_0(x)$] and $G_a(x)$ (since U_1 and U_2 are equal in area).

to make comparisons across conditions, across modalities, across occasions, and across tasks for a single subject and also between subjects, independently of the physical scales or even the modalities used.

If $\mu_n = 0$ and $\sigma_n = 1$, $d' = \mu_s$. However, this does not fully represent the relation between the distributions; we also need to know the parameter σ_s . Thus, for $f_n(x)$ and $f_s(x)$ in Figure 1, Equation 2 gives $d' = a/(a/2) = 2$. But d' then has the same value whether we compare $f_n(x)$ and $f_s(x)$ or $f_n(x)$ and $f'_s(x)$. Yet the latter pair of density functions overlap more extensively and thus must give lower sensitivity. We distinguish between the two cases by noting that, in the first, $\sigma_s/\sigma_n = 1$ and, in the second, $\sigma_s/\sigma_n = 2^{1/2}$.

The information provided by the noise and signal density functions is equally well represented by the noise and the signal distribution functions, $F_n(x)$ and $F'_s(x)$ [or $F'_s(x)$], as is illustrated in Figure 1A. μ_{dif} , the numerator of d' , corresponds to the lateral separation of the two distribution functions. This may be measured at any probability level when $\sigma_s = \sigma_n$ but must be taken at the 50 percent points when $\sigma_s \neq \sigma_n$. It will now be useful to establish a relation between the area between the distribution functions and μ_{dif} .

μ_{dif} and the Area Between the Distribution Functions

Let us call the area contained between the two distribution functions, $F_n(x)$ and $F_s(x)$, α_b . Evidently, as d' increases, α_b will increase accordingly. Figure 2 illustrates the simple case in which the two probability distribution functions are step functions, $S_0(x)$ with a mean of 0 ($S_0(x) = \{0, x < 0; 1, x \geq 0\}$), represented by the vertical line $0c$, and $S_a(x)$ with a mean of a ($S_a(x) = \{0, x < a; 1,$

$x \geq a\}$), represented by ab . Evidently the area between these two distribution functions is the area of the rectangle $0abc = a = \mu_{\text{dif}}$. If we substitute for $S_0(x)$ the ramp function $L_0(x)$, also with a mean of 0, the area between the curve centered on 0 and $S_a(x)$ is unchanged: α_b is reduced by the area of the triangle T_1 and is increased by the area of the triangle T_2 , and these triangles are equal in area. Thus, the relation $\alpha_b = \mu_{\text{dif}}$ holds for such functions. A ramp function, not necessarily of the same slope, can similarly be substituted for $S_a(x)$ without altering α_b .

Figure 2 also shows $G_a(x)$, a monotonic distribution function with odd symmetry and a mean of a . If we substitute $G_a(x)$ for $S_a(x)$, it appears that α_b will be effectively unchanged, since it will be increased by the area U_1 and decreased by the area U_2 , which are the same by symmetry. A corresponding argument would apply if a similar substitution were made for $S_0(x)$.

We can prove that $\alpha_b = \mu_{\text{dif}}$ in the more general case as follows. Consider two parallel monotonic continuous probability distribution functions, $G_0(x)$ and $G_a(x)$ with means of 0 and a [e.g., $F_n(x)$ and $F_s(x)$ in Figure 1A]. Then,

$$\begin{aligned}\alpha_b &= \int_{-\infty}^{\infty} [G_0(x) - G_a(x)] dx \\ &= \int_{-\infty}^{\infty} [G_0(x) - G_0(x-a)] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^x g_0(t) dt - \int_{-\infty}^{x-a} g_0(t) dt \right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{x-a}^x g_0(t) dt \right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_t^{t+a} g_0(t) dx \right] dt \\ &= \int_{-\infty}^{\infty} g_0(t) [a] dt = a.\end{aligned}$$

Thus,

$$\alpha_b = \mu_{\text{dif}} = \mu_s - \mu_n. \quad (3)$$

This assumes that $G_0(x)$ and $G_a(x)$ have the same slope. However, if we refer to the distribution functions $F_n(x)$ and $F_s(x)$ in Figure 1A, we may note that if we substitute $F'_s(x)$ for $F_s(x)$, the effect is to add to α_b the area above the 50 percent point,

$$\alpha_u = \int_a^{\infty} [F_s(x) - F'_s(x)] dx,$$

and remove the area below the 50 percent point,

$$\alpha_l = \int_{-\infty}^a [F'_s(x) - F_s(x)] dx,$$

and these areas are equal. Thus, we can apply Equation 3 to pairs of parallel continuous functions with odd symmetry about the 50 percent point and also to such functions that differ in slope, provided any intersection between the two functions occurs negligibly close to 0 or 1.

Balakrishnan's (1998b) New Measure of Sensitivity, S'

Equation 3 suggests that if we could estimate the area between the two probability distribution functions underlying a discrimination experiment, this would estimate

Figure 3. Panel B shows density functions for noise and signal $f_n(x)$ and $f_s(x)$, with means of $\mu_n = 0$ and $\mu_s = a$ and standard deviations of $\sigma_n = \sigma_s = a$. The decision criterion, $x_c = x_5 = a/2$, determines whether the response is yes or no. The five levels of confidence for each response are determined by criteria $x_1 = -(3/2)a, x_2 = -a, \dots, x_{n-1} = 2.5a$, giving $n = 10$ response categories. Panel A shows the corresponding cumulative rating responses $F_n(R \leq k)$ and $F_s(R \leq k)$ as histograms plotted against the decision axis.

$\alpha_b = \mu_{\text{dif}} = \mu_s - \mu_n = a$ and thus provide a measure of the separation between the two distributions. This is what Balakrishnan (1998b) appears to be doing with his new sensitivity index, S' , which he puts forward as “an estimate of the difference between the means of two distributions” (p. 69).

Balakrishnan (1998b) derives S' for the rating procedure only. The derivation, slightly simplified, is illustrated in Figure 3. (Balakrishnan has an extra transformation, from sensory input to “subjective likelihood ratio.” Since this has no experimental implications in the present context, it is omitted here.) Panel B shows density functions for noise $f_n(x)$ and signal $f_s(x)$, with means of $\mu_n = 0$ and $\mu_s = a$, and standard deviations of $\sigma_n = \sigma_s = a$ (giving $d' = 1$). Rating criteria are also shown. Following Treis-

man and Faulkner (1984b), we refer to these as the decision criterion and the confidence criteria. The decision criterion, here $x_c = x_5 = a/2$, determines the direction of response, yes or no. The figure shows $n = 10$ rating responses, corresponding to five levels of confidence for yes and five for no, “5” and “6” responses being the least confident and “1” and “10” the most confident. These responses are determined by confidence criteria (the vertical dashed lines) $x_1 = -(3/2)a, x_2 = -a, \dots, x_k, \dots, x_{n-1} = 2.5a$. In this illustration, the confidence criteria are equally spaced. In analyzing rating data, the responses in each category for a given stimulus can be cumulated to give a cumulative response histogram. Such histograms, $F_n(R \leq k)$ and $F_s(R \leq k)$, where R is the response and k is the category index, are plotted against the decision axis scale in Figure 3A, as they might be if we had access to that scale. They can be considered as approximations to $F_n(x)$ and $F_s(x)$. The difference between the cumulated probabilities at each criterion k , $F_n(R \leq k) - F_s(R \leq k)$, is shown as a dotted vertical line for that value of the criterion, $1 \leq k \leq n - 1$.

S' is derived as follows. First, for each rating category k , Balakrishnan (1998b) defines the quantity

$$U_R(k) = F_n(R \leq k) - F_s(R \leq k). \quad (4)$$

Then, “[a]ssuming that the spacing between the criteria ... is constant” (Balakrishnan, 1998b, p. 72) and that the interval $x_k - x_{k-1} = 1$, the area under $U_R(k)$ is given by

$$S' = \sum_{k=1}^n F_n(R \leq k) - F_s(R \leq k) = \sum_{k=1}^n U_R(k). \quad (5)$$

Before comparing S' and d' , we will consider some features of S' .

1. What does S' measure? For the 10 categories in Figure 3A, the sum of the vertical differences (dotted lines) gives $S' = 1.94$. This is not a , the difference between the means, which Balakrishnan (1998b) says it estimates. It is a sum of probability differences taken at those points at which the subject chooses to place a criterion. It becomes the estimate of an area under a curve when the interval between neighboring criteria is known. Balakrishnan defines these intervals as unity, which defines an arbitrary scale for the decision axis. But we do not know the relation between this scale and the decision axis scale x on which $\mu_s = a$, so we cannot relate one to the other. In Figure 3, the spacing between the criteria is $0.5a$. If we substitute this value for unity, the inter-criterion difference assumed by Balakrishnan in calculating S' , we get $S'(0.5a) = 0.97a$, which is a better estimate of a . But to obtain this estimate, it was necessary to know the inter-criterion interval in terms of the value of a , the value which S' is intended to estimate! Without such prior knowledge, we could not get from $S' = 1.94$ to an estimate of a . Nor can we use S' to estimate the standardized difference between the means, here $d' = 1$.

2. A second problem in interpreting the quantitative value given by S' is that the value obtained depends on the number of criteria specified by the experimenter. If the

number of categories in Figure 3 were doubled (and fell in the range shown), twice as many differences $F_n(R \leq k) - F_s(R \leq k)$, of similar values to those shown in Figure 3A, would be summed to give S' , roughly doubling its value. If the number of criteria were halved, the value of S' would be roughly halved.

3. The experimenter may specify n categories, but this does not ensure that the subject uses n categories: Some responses may not be employed. But the value of S' depends on the number of categories actually used. For example, if the subject always used the same level of confidence—for example, “4” for *no* responses and “7” for *yes* responses—only one criterion, x_c , would be in play, and for the data in Figure 3, this would give $S' = 0.383$.

4. S' is also affected by where the subject places the criteria he or she does employ. For example, if a single criterion were employed and were placed at $-a$, we would have $S' = 0.136$.

5. Figure 3 is a favorable case for Balakrishnan's arguments, since we show the criteria as equally spaced. We have no right to assume that this would hold in any real case, since the actual spacing of the criteria depends on where the subject chooses to place them and equal spacing is likely to be rare. (This is a point worth stressing. Typically, confidence criteria in the rating procedure have been found to be unequally spaced [e.g., Egan, Greenberg, & Schulman, 1961; Egan, Schulman, & Greenberg, 1959; Stretch & Wixted, 1998; Treisman & Faulkner, 1984a, 1985; Treisman & Williams, 1984], as can frequently be seen by simple inspection of standardized rating procedure ROCs.) When equal spacing does not hold, Balakrishnan's assumption that the spacing between the criteria is constant (and equal to unity) makes the assessment of S' even more problematic. Suppose the subject employs four criteria and these are actually placed at 0, 0.5, 2, and 2.25 on the decision axis. Equating the three intercriterion intervals to unity imposes a series of expansions and contractions on the decision axis scale, whose effect is to base S' on an unknown nonlinear transformation of this scale. This transformation of the underlying scale is not only undetermined, but also unstable, since it will be different for every subject's and session's freshly chosen criterion placements.

It may appear that these considerations cast a shadow over S' and that no weight deserves to be given to numerical values of this statistic. We should note that Balakrishnan (1998b) is not unaware of these difficulties and does not claim that numerical values of S' are meaningful. But he does not regard this as a difficulty. He believes that “ordinal relationships between conditions tend to be more important than absolute levels of performance in psychological research” (p. 69) and argues that “like d' , the numerical value of S' should not be used as an estimate of the true spacing between the distributions, but instead as an index of relative performance level” (p. 72).

The comparison with d' is misleading. It is, indeed, regarded as an estimate of the normalized distance between noise and signal distributions and, for this reason, is used as a measure of sensitivity. We can, of course,

calculate d' as simply an index of performance where we do not have a model of the underlying processes. But d' can also be embedded in a theoretical context in which it estimates a parameter of a model, since SDT allows us to frame models of underlying processes and estimate their parameters. An advantage of this approach to understanding psychological processes is that it allows predictions to be made and tested and it allows us to relate the results of different types of psychophysical procedures (Swets, 1964; Treisman, 1999; Treisman & Watts, 1966). S' cannot serve these purposes.

The question that arises at this point is the following: Under what circumstances might S' be more useful than d' ? Balakrishnan argues that S' can be used to compare relative performance in different conditions. But in general, S' will not be very useful as an index of relative performance. If, for example, we wish to compare the performance of 2 subjects run in the same experimental condition, the two values of S' will be vitiated to an unknown extent by the different choices the 2 subjects make in placing their criteria and how many they each employ. If we wish to compare the same subject's performance when, say, sober and when affected by alcohol, the same problem of the effects of uncontrolled criterion placement will arise. In contrast, d' is not affected by the placement of criteria.

Evidently, for S' to be useful in comparing two experimental conditions, the experiment must be so designed so as to make it likely that the criteria will be identically, or at least very similarly, placed in the two conditions. Consider a rating procedure design with a *low* condition in which a weak stimulus s_L must be distinguished from noise n and a *high* condition in which a slightly stronger stimulus s_H must be distinguished from noise. If s_L and s_H are themselves difficult to discriminate and if trials of the two conditions are intermingled in each block, s_L and s_H will be tested against the same criteria. In that case, the estimate of $U_R(k)$ for each criterion should be slightly greater for the *high* condition, giving a larger value of S' . A similar effect may be obtained if the *low* and the *high* conditions are similar and are given in separate small blocks that are intermingled in a session and the subjects are well practiced.

Balakrishnan (1998b) has conducted simulations of such designs with artificial data. In a typical case, the rating procedure with 10 categories was used to derive S' and other sensitivity measures, including different d' estimates, for discrimination between a noise distribution with $\mu_n = 1.0$ and a low-valued signal s_L with $\mu_s = 1.1$ ($\sigma_s = \sigma_n = 1$). This calculation was repeated for the same noise distribution and a higher signal s_H with $\mu_s = 1.2$. Criterion placements were varied between comparisons but appear to have been the same within a comparison. The author asks the following question: How often do the values obtained for S' and for d' return a lower value for the s_L data than for the s_H data? Results obtained from simulations with sample sizes of 25–100 samples per run (Balakrishnan, 1998b, Table 2) provided evidence that, in this case, the probability of misordering s_L and s_H was

less for S' than for the other measures examined. Calculations were also performed using small samples (10–50) resampled from experimental data sets. In these data, s_L and s_H were very similar (e.g., lines of length 6.75 and 6.9 cm, with $n = 6.6$ cm), and the two conditions were presented in intermingled short blocks in 2-h sessions with well-practiced subjects (Balakrishnan & Ratcliff, 1996). The previous finding was supported (Balakrishnan, 1998b, Table 3).

In no cases were the differences between the different measures tested statistically. In comparing two statistics, one would normally consider their means and standard errors, how accurate they are as estimators and whether biased or unbiased, and what it is they estimate. But we cannot easily do this with S' , since it may not be clear what a given value of S' can be taken to estimate and what would be an accurate result. Instead, the author relied on the use of large numbers of trials.

If the estimation of d' from rating data involves curve-fitting, this may be unsatisfactory if there are many categories and sample sizes are too small (McNicol, 1972): Not all the categories may be employed, categories may have to be collapsed in analysis, it may not be possible to estimate σ_s/σ_n , and outlier values of d' are readily produced. Balakrishnan (1998b) used 10 categories, with the sample sizes given above.

In conclusion, the question is the following: When are d' or S' likely to be more useful? A large number of SDT procedures have been developed that allow d' to be estimated, and this measure has been widely applied (Green & Swets, 1966; Macmillan & Creelman, 1991). Of these procedures, S' can be evaluated only for the rating procedure. In some applications, we may be interested in the actual value of the sensitivity measure; for example, $d' = 1$ may be used to define a threshold. In such cases, S' is of no use. We may be interested in comparing sensitivity performance in two conditions—for example, for different subjects doing the same task, for the same subject in different physiological states, or for different types of stimulation. In such cases, we cannot assume that criteria are placed identically or similarly in the two conditions, and S' cannot be relied on. However, Balakrishnan's (1998b) work leads to the valuable observation that for rating procedure experiments in which the criteria are likely to be substantially unchanged between two conditions, and if our only interest is in ordering the two conditions, S' may have an advantage over d' and a number of other measures.

Such experimental designs have been discussed above. They involve presenting two signal strengths, s_L and s_H , and using the subject's performance to determine which is the stronger stimulus. This is not a question we commonly ask, because we can usually measure the intensities or magnitudes of the stimuli we employ directly. But the question might arise if we wished to determine the relative discriminability of complex stimuli that do not have an obvious physical ordering—for example, pairs of words or visual patterns. Balakrishnan's evidence suggests that

S' might be useful in such a case. However, this observation does not justify claiming that, in general, S' is "better justified and statistically more reliable than . . . d' " (Balakrishnan, 1998b, p. 69).

New Measures of Response Bias: SDT Thrice Refuted?

Balakrishnan (1998a, 1998b, 1999) has developed three new related measures or tests that he uses to examine whether response bias occurs in discrimination, using the rating procedure. He claims that these show that SDT is fundamentally flawed: Its central assumption, that manipulating payoffs or probabilities can shift the decision criterion, is invalid. When his measures are used, the criterion does not shift. Even though he also finds that changes in observed β values, hit rates, and false positives all occur as predicted by SDT, he argues that since SDT fails, such observations require a new explanation. If Balakrishnan is correct, he has brought into question half a century's work in psychophysics. I shall examine his three methods in turn.

The $U_R(k)$ curve. Figure 3A illustrates the cumulated response histograms $F_n(R \leq k)$ and $F_s(R \leq k)$ predicted by the SDT model in Figure 3B. They can be regarded as estimates of $F_n(x)$ and $F_s(x)$. The histograms are plotted against the decision axis, on the assumption that the criteria are equally spaced. For each criterion, Equation 4 defines $U_R(k)$ as the difference $F_n(R \leq k) - F_s(R \leq k)$; these values correspond to the vertical dotted lines in Figure 3A. In Figure 4, panels A and B illustrate the SDT analysis of a discrimination or detection task with equally spaced criteria and with the decision criterion $x_c = x_5$ placed at $0.5a$, the unbiased position for which $\beta = l(x_c) = 1$. The nine confidence criteria x_k range from $x_1 = -1.5a$ to $x_9 = 2.5a$, at intervals of $0.5a$. In Figure 4A, the values of $F_n(x) - F_s(x)$ derived from these density functions are plotted against the decision axis. As $F_n(x)$ grows more rapidly than $F_s(x)$ below $l(x) = 1$ and more slowly above, the peak of the $F_n(x) - F_s(x)$ function here coincides with x_c .

The models in Figures 4C and 4D differ only in that the decision criterion has shifted to $x_c = x_5 = 1.5a$. The confidence criteria have moved to the right in concert with x_c , as it has been shown confidence criteria do (Treisman & Faulkner, 1984b), and the spacing is unchanged at $0.5a$. The decision criterion is now "biased" in that $l(x_c) \neq 1$; here, it is $l(x_c) = 2.38$.

Figures 4A and 4C illustrate that when the criterion is unbiased, the peak of the $F_n(x) - F_s(x)$ function coincides with x_c . But when the criterion is biased, the peak of the $F_n(x) - F_s(x)$ function is displaced from x_c . Thus, in Figure 4C, x_c has moved to $1.5a$, but the peak of the function remains at the unbiased position $0.5a$ (which now coincides with x_3). These are the SDT predictions that Balakrishnan sets out to test. The argument is that $U_R(k)$, obtained from a rating procedure, estimates the curve $F_n(x) - F_s(x)$. Thus, when x_c is unbiased (Figures 4A and 4B), the peak of the $U_R(k)$ curve should correspond to the decision criterion $x_c = x_5$, with positive slope to the left of

Figure 4. Panel B shows noise and signal distributions with the same parameters as those in Figure 3B. The decision criterion $x_c = x_s = 0.5a$ has $\beta = I(x_c) = 1$. In panel D, the decision criterion is shifted to $x_c = x_s = 1.5a$, giving $\beta = I(x_c) = 2.38$. In each case, the confidence criteria are spaced at $0.5a$ intervals with $x_5 = x_c$. Panels A and C show the corresponding values of $F_n(x) - F_s(x)$ plotted as continuous curves against the decision axis.

x_c and negative slope to its right. Such an experimental finding would be consistent with an unbiased decision criterion. But if the criterion can be made to shift—for example, to the right, as in Figures 4C and 4D—the boundary between *no* responses and *yes* responses will lie to the right of the $U_R(k)$ peak. Between the peak (at x_3 in Figure 4C) and the decision criterion (at $x_c = x_5$), the slope of the curve will be negative. Thus, a biased criterion (x_c shifted to the right) is indicated by a negative slope to the left of x_c or (if x_c is shifted to the left) by a positive slope to the right of x_c . This observation is the foundation for Balakrishnan's (1998a, 1998b, 1999) first test of bias. In experiments in which a payoff or frequency manipulation was used that, according to SDT, should shift the criterion to a biased position, the author claims that displacements of x_c from the $U_R(k)$ peak did not occur. Thus, these manipulations do not shift the criterion, and SDT is refuted.

However, there are reasons to be cautious about these claims. They depend on the use of the experimental cumulated response histograms $U_R(k)$ to approximate the

theoretical $F_n(x) - F_s(x)$ curve. But although the latter is shown in the model plotted against the decision axis x , it is not possible to plot $U_R(k)$ against x . To do that would require that we know or estimate the positions of the criteria on x , but that cannot be done using Balakrishnan's distribution-free approach. What the author does instead is to plot $U_R(k)$ against criterion number, $1, \dots, k, \dots, n-1$. But if we do not know the criteria to be equally spaced (and we do not), the scale given by plotting criterion numbers at intervals of unity is not the decision axis scale but an unknown, possibly nonlinear, transformation of that scale. (Especially so since Balakrishnan's experimental instructions are intended to produce unequal intercriterion spacings and appear to succeed in this.) Plotted on the criterion number scale, $U_R(k)$ will then estimate not the $F_n(x) - F_s(x)$ function shown in Figure 4, but a distorted version of that function. Because of this, we cannot rely on the model's features, such as the reversal of slope between x_3 and x_c in Figure 4C, being clearly reflected in the $U_R(k)$ plot. Thus, we need to consider carefully the effects of the scale transformation on the SDT predictions in Figure 4

and examine whether, when these are taken into account, Balakrishnan's $U_R(k)$ curves present as big a challenge for SDT as he believes.

It turns out that predictions that look clear enough when we consider the theoretical smooth curves and equally spaced criteria in Figure 4 may look different when the curve is approximated by rating data and the criteria are not equally spaced. Figure 5 shows the SDT predictions for $U_R(k)$ for three cases. Panels B, D, and F show the same theoretical noise and signal distributions, with $\mu_n = 0$, $\mu_s = 1$, $\sigma_s = \sigma_n = 1$. In Figure 5B, the criterion is unbiased, $\beta = l(x_c = 0.5) = 1$. In Figures 5D and 5F, the criterion is $x_c = 0.9$, giving $\beta = l(x_c = 0.9) \approx 1.5$, the value used in Balakrishnan's (1998b) experiment. In his experiments, subjects were instructed "to be cautious about using the extremes of the rating scale" (Balakrishnan, 1998b, p. 82). To accommodate this, most intercriteria spacings in Figure 5 are larger than those in Figure 4, so that Criteria 1 and 9 move outward, reducing the probabilities of the most confident responses, "1" and "10." Also, in Figures 5B and 5D, Criteria 4 and 6 lie 0.1 units below or above x_c , to reduce production of the least confident responses, "5" and "6." In Figure 5F, Criteria 4, 5, 6, and 7 share the same location, so that the "5," "6," and "7" responses do not occur at all and the spacing of criteria 8 and 9 is correspondingly increased.

Figures 5A, 5C, and 5E show the values of $U_R(k)$ that would be predicted by each of these SDT models, plotted against criterion number.

Figure 5B has $l(x_c) = 1$, as would be expected when there are no asymmetric payoffs or probabilities. This

predicts that the peak of the $U_R(k)$ function will lie at the decision criterion, as is shown in Figure 5A. However, the nonlinear distortion entailed when criterion number is substituted for the decision axis has required Criteria 4, 5, and 6 to be plotted at the same unit intervals as the other criteria. Thus, the narrow band between x_4 and x_6 on the decision axis broadens out to give a flat-topped curve in Figure 5A. There is a peak at x_5 , but it is not visually distinguishable from the values at x_4 and x_6 .

In Figures 5C and 5D, the decision criterion has shifted to the position that gives $l(x_c) = 1.5$, with the confidence criteria shifting similarly, and the least confident responses, "5" and "6," are again rare. According to Figure 4C, when the criterion is displaced to the right of $\beta = 1$, the peak of the $F_n(x) - F_s(x)$ function should lie to the left of the decision criterion. Figures 5C and 5D accord with this. The peak of the $U_R(k)$ function occurs at x_4 , to the left of x_c , as can just be discerned. Balakrishnan (1998b, p. 84) says of his data, "clearly, none of the empirical functions . . . suggest a strong response bias, because none of the modes occur sharply to one side of the rating scale." This contributes to his claim to refute SDT. This argument relies on the shape of the $U_R(k)$ function. But Figure 5C illustrates that the lack of such a sharp differentiation is just what is predicted by applying Balakrishnan's procedures to the SDT model.

The model in Figure 5F also has the decision criterion at the location that gives $l(x_c) \approx 1.5$, but here the simulated subject has responded to Balakrishnan's instructions by avoiding the responses "5," "6," and "7" completely, so that Criteria 4, 5, 6, and 7 converge on x_c , as shown. The

Figure 5. Panels B, D, and F show noise and signal distributions, with means of $\mu_n = 0$ and $\mu_s = 1$ and standard deviations of $\sigma_s = \sigma_n = 1$. $P(s) = P(n) = .5$. In panel B, the decision criterion, $x_c = x_5 = .5$, giving $\beta = l(x_c) = 1$. In panels D and F, the criterion is $x_c = x_5 = 0.9$. The spacing of the criteria is described in the text. Panels A, C, and E show the values of $U_R(k)$ predicted by each model, plotted against criterion number.

remaining criteria to the right of x_c are accordingly more widely spaced, to cover the same range. The predicted $U_R(k)$ curve in Figure 5E, plotted against criterion number, appears to be roughly centered on the decision criterion, and the highest value of $U_R(k)$ is again found at x_4 , as in Figure 5C, but also at x_c , x_6 , and x_7 , rendering the peak a flat-topped plateau extending over this range. Thus, even though the peak of the theoretical $F_n(x) - F_s(x)$ curve lies to the left of x_c (i.e., the decision criterion is biased toward *no* responses), interpolation would place the peak of the derived $U_R(k)$ function at 5.5, to the right of x_c (indicating a bias toward *yes* responses).

When Figures 5A, 5C, and 5E are compared, it is apparent that in Balakrishnan's procedures, the appearance of the $U_R(k)$ curve is largely dependent on the details (unobservable in the distribution-free approach) of how subjects place their criteria. This can easily lead to curves whose shapes are at variance with Figure 4. Such results cannot constitute evidence for rejecting SDT.

Balakrishnan is aware that if criterion spacings about the decision criterion are broad, his procedure becomes very insensitive. (For example, in Figure 5D, the density functions intersect between x_3 and x_4 , giving a positive $F_n(x) - F_s(x)$ slope to the left of 0.5 and a negative slope to its right, but $U_R(k)$ cannot capture this: It must place the peak either at x_3 , with a negative slope to its right, or at x_4 , with a positive slope to its left.) To avoid this, Balakrishnan attempts to ensure narrow criterion spacings in the vicinity of x_c by instructing his subjects to avoid the extreme responses ("1" and "10," and "5" and "6"). The reduced use of the low confidence responses will lead to the narrow spacings between criteria x_4 , x_5 , and x_6 illustrated in Figures 5B and 5D, and if the least confident responses are avoided altogether, the criteria would merge, as in Figure 5F.

Unfortunately, this procedure is self-defeating. If the least confident responses are not used at all, as in Figure 5F, we are back with broad categories bounding x_c and an insensitive procedure. If the subject does use the "5" and "6" responses, but rarely, as in Figure 5D, we have the desired narrow category spacings about x_c . But this entails that we have only small samples of responses in these categories. Consider the effects of these small samples on accuracy. In Figure 5D, the proportion of the *s* density function lying between x_4 and $x_5 = x_c$ is greater than the proportion of the *n* density function in this range, and so the slope of $F_n(x) - F_s(x)$ should be negative between these two criteria, as is shown by the theoretical $U_R(k)$ curve in Figure 5C. But these two category boundaries are close together precisely because Balakrishnan's instructions have ensured that the sample of "5" responses is very small. But observations based on small samples have large standard errors. With a sample of only one or a few "5" responses, the probability in an actual experiment that the number deriving from the *n* distribution might randomly exceed the number deriving from *s* and, thus, give a *positive* slope from 4 to 5 and an "unbiased" peak at x_c could

be quite high. Evidently, detecting the modes of these flat-topped $U_R(k)$ functions is subject to a degree of error that has not been estimated or allowed for.

Inspection of Balakrishnan's (1998a, 1998b, 1999) $U_R(k)$ curves confirms that they are generally flat-topped, much like the SDT predictions in Figure 5. Nevertheless, from such evidence, Balakrishnan (1999) has concluded that the decision rule in his experiments was uniformly unbiased despite payoff and base rate manipulations. He concludes that "this seemingly paradoxical result is fundamentally inconsistent with the entire family of signal detecton theory models, raising some important questions about the significance of many published results" (Balakrishnan, 1999, p. 1189). The logic of this argument is that of accepting the null hypothesis. On the present analysis, we have to conclude that the power of a procedure in which $U_R(k)$ values are plotted against criterion number and the position of the mode is assessed by inspection is likely to be so low that the belief that such data refute SDT cannot be seriously entertained.

A new index of response bias: Ω_p . The second test believed by Balakrishnan (1998a, 1999) to refute SDT is a new measure of response bias Ω_p . He claims that when SDT predicts that the decision criterion should shift, his new measure shows that little or no bias is produced: The criterion does not shift. In a series of experiments (Balakrishnan, 1998a, 1999) in which payoffs or signal probabilities were varied, he observed the shifts in β and in the hit and false positive rates predicted by SDT, but he found that Ω_p gave negligible or zero values. This showed that no bias was induced by payoff or probability manipulations and that shifts in β and in false positive rates require some new explanation. This surprising result, if valid, would require a fundamental reappraisal of SDT. What weight does it deserve?

Balakrishnan's Ω_p is closely related to the $U_R(k)$ test. It is an attempt to give it quantitative form. He defines Ω_p as the proportion of responses that arise in the region between the value of x at which $l(x) = 1$ and the value x_c of the decision criterion, and he uses the slope of the $U_R(k)$ curve to identify this proportion. Referring again to Figure 4, in panel A the slope of the curve is positive for all *no* responses and negative for all *yes* responses, the criterion has $l(x_c) = 1$, and there are no biased responses: $\Omega_p = 0$. But in panel D, the decision criterion is biased, giving $l(x_c) > 1$. The density functions intersect at x_3 , giving $l(x_3) = 1$. In the region between x_3 and x_c , "biased" responses occur: The *no* response is given, although $f_n(x) < f_s(x)$ over this range. Consequently, between x_3 and the decision criterion x_5 in Figure 4C, the slope of the $F_n(x) - F_s(x)$ curve is negative. Elsewhere, the responses are unbiased.

This negative slope to the left of the decision criterion reveals that the corresponding responses (in Figure 4C, these are the "4" and "5" responses) are biased. Ω_p is intended to estimate their overall proportion as a measure of bias. Ω_p is defined as the proportion of responses for which $U_R(k)$ has a negative slope below x_c , a positive

slope above x_c , or both. Let us examine how well this procedure works for the theoretical illustrations in Figure 5.

In Figures 5A and 5B, the criterion is defined by $l(x_c) = 1$. Thus, there are no biased responses, and $U_R(k)$ should increase below x_c and decrease above it, as in Figure 4A. In Figure 5A this is seen, and so Balakrishnan's procedure gives us $\Omega_p = 0$, which is exactly right.

In the model in Figure 5D, the decision criterion is shifted from $x = 0.5$ to $x = 0.9$. This gives $l(x_c) \approx 1.5$, the value in Balakrishnan's (1998b) experiment. This shift causes biased responses to occur between $x = 0.5$ and $x_c = 0.9$. The proportion of biased responses predicted by SDT for these parameters is thus the mean proportion of the two density functions lying between $x = 0.5$, where the functions intersect, and $x_c = 0.9$. This is .158.

A valid measure of bias intended to determine the proportion of biased responses should approximate this value. However, when this proportion has to be estimated from the rating procedure, there is a problem. In Figure 5D, biased responses occur between the intersection of the density functions at $x = 0.5$ and $x_c = 0.9$, but there is no criterion at $x = 0.5$. The bounding criteria are $x_3 = 0$ and $x_4 = 0.8$. But in Figure 5C, the slope of the $U_R(k)$ curve is positive between x_3 and x_4 : Overall, more *no* responses are added than *yes* responses over this range. Following Balakrishnan's procedure, Ω_p would be derived as the proportion of "5" responses, since between $k = 4$ and $k = 5$ on the $U_R(k)$ plot, the curve has a negative slope. This proportion is $\Omega_p = .034$. This is less than a quarter of the true proportion of biased responses, which is .158.

However, Balakrishnan has also defined an upper bound for Ω_p , to deal with cases in which it might produce an underestimate. He advises that "when Ω_p is zero, the researcher should also calculate the proportion of least confident no-signal responses and the proportion of least confident signal responses. The maximum of these two values is an upper bound on the true proportion of biased responses" (Balakrishnan, 1998a, p. 608). He also calculates this upper bound for small values of Ω_p .

To calculate this upper bound for Figure 5C, we need also to find the proportion of "6" responses. This is .033. We already have the proportion of "5" responses, .034. The greater of the two values is the *upper bound*: This is again .034. So, for Figure 5C, we have an "upper bound" for the proportion of biased responses that is little more than one fifth of the real value.

For the model in Figure 5F, the parameters are the same, so the true proportion of biased responses is again .158. In Figure 5E, the slope of the $U_R(k)$ curve is nowhere negative below x_c and nowhere positive above it. No responses show a reversal of slope, and so $\Omega_p = 0$. We again turn to calculate the upper bound. This is the greater of the proportion of "5" responses or "6" responses. However, there were no "5" responses or "6" responses, so the upper bound is again zero. This is an even grosser underestimate of the true value, .158.

These two calculations show that both Ω_p and its upper bound, as defined by Balakrishnan (1998a), may grossly underestimate the true proportion of biased responses.

The author claims that his low values of Ω_p are striking evidence that the decision criterion does not move when SDT predicts it should and shifts in β apparently show that it has. The values of Ω_p for 4 subjects in his asymmetric payoff condition range from 0 to .006 (Balakrishnan, 1998b, p. 84), indicating no bias, whereas the corresponding mean values of $P(\text{NO} | n)$ and $P(\text{YES} | s)$ are .81 and .67, clearly different. This difference in response probability is in the same direction for each subject and is statistically significant ($p < .001$) for each. For SDT, this is direct evidence that the payoff manipulation has shifted the criterion. In view of their potential for underestimation, it does not appear that Balakrishnan's low values of Ω_p are capable of refuting this.

I have not stressed, but it should be obvious, that the calculation of Ω_p also suffers from the difficulty in determining the position of the $U_R(k)$ mode discussed above. This is not a point that Balakrishnan accepts: He says of the main experiment in Balakrishnan (1998a, p. 612), "the sample sizes were large enough for the estimated functions to be smooth and unimodal (i.e., the result cannot be attributed to estimation error)." This belief—that if the number of trials is large, the position of the mode should be clearly evident—may explain why he does not address the difficulty of specifying exactly where the mode is with his procedure, nor does he take the precaution of determining a standard error for estimates of the mode and for measures that rely on the position of the mode.

In fact, the position of the mode may be greatly affected by a small subset of trials, however large their total number may be. For example, in the model in Figures 5C and 5D, we have seen that the expected proportion of "5" responses is .034. Thus, if in an experiment with these parameters there were 100 signal and 100 noise trials, we might have seven "5" responses in total. The probability that four or more of these might randomly be generated on noise trials and three or less on signal trials would be quite high. Such an outcome would give a positive slope between $k = 4$ and $k = 5$, causing the mode to appear at x_c . Thus, the fact that the total number of trials might be 200 would not guarantee that the position of the mode is not subject to estimation error, and this might be quite large.

A third test: Detection of suboptimality. Balakrishnan (1998a) also relies on a third test for the placement of the criterion. This is based on a consideration of the effects of criterion location on the probability of a correct response for each response category, $P(C | R = k)$. To illustrate this, Figure 6 gives SDT predictions for three cases. Panels B, D, and F show the density functions on the decision axis for normal distributions with $\mu_n = 0$, $\mu_s = 2$, and $\sigma_s = \sigma_n = 1$, giving $d' = 2$. The probability of the signal is $P(s) = .1$ in each case, as in Balakrishnan's (1998a) experiments. Balakrishnan's (1998a) subjects were instructed to be conservative in using the extreme confidence categories, and the flat-topped $U_R(k)$ curves obtained suggest that they obeyed this instruction. Thus, the spacing between criterion positions x_4 , x_5 , and x_6 shown in these models is small. Figure 6B models the case in which the criterion remains unbiased despite asymmetric probabili-

Figure 6. Effects of three criterion positions on $P(C|R=k)$. Panels B, D, and F show noise and signal distributions with $d' = 2$ and $P(s) = .1$. In panel B, $x_c = 1, l(x_c) = 1$. In panel D, $x_c = 1.8, l(x_c) = 5$. In panel F, $x_c = 2.1, l(x_c) = 9$. Nine criterion values are shown in each case. The inter-criterion spacing is 0.8 units, except between x_4, x_5 and x_6, x_7 , where it is 0.1 unit. Panels A, C, and E plot the corresponding values of $P(C|R=k)$.

ties: $l(x_c = 1) = 1$. In contrast, in Figure 6F, the decision criterion is located at the optimal position prescribed by SDT for the signal probability $P(s) = .1$. This criterion has the value $x_c = 2.1$, which gives $l(x_c) = 9$, the optimal value. In Figure 6D the criterion, $x_c = 1.8$, is intermediate in position, giving $l(x_c) = 5$. It has shifted in response to the probability manipulation, but not to the full theoretical extent. SDT studies have generally found that manipulations intended to shift the criterion succeed in doing so, but not usually to the full extent expected (e.g., Green, 1960; Green & Swets, 1966). (For reasons that shifts may be incomplete, see Treisman & Williams', 1984, criterion setting theory.) Thus, Figure 6D represents the result most strongly suggested by classical studies.

For Categories 1–5 (on the 10-point scale), the probability of a correct response for a given category is the probability that the sensory inputs falling in the domain on the decision axis corresponding to that category arise from noise; for Categories 6–10, it is the probability that they arise from the signal. Thus, the probability of a correct response for $k \leq 5$ is

$$P(C|R=k) = \frac{P(n) \int_{x_{k-1}}^{x_k} f_n(x) dx}{\left\{ P(n) \int_{x_{k-1}}^{x_k} f_n(x) dx + P(s) \int_{x_{k-1}}^{x_k} f_s(x) dx \right\}} \quad (6)$$

(with the lower bound at $-\infty$ for $k = 1$). A similar expression applies for $k > 5$.

If the criterion is located at the optimal position for a given $P(s)$, with homogenous payoff values, $P(C|R=k)$ must be .5 or greater for every response category. In Figure 6F, the criterion is at the optimal position for $P(s) = .1$, and in Figure 6E, the corresponding values of $P(C|R=k)$ are all greater than or equal to .5. This is the observation Balakrishnan (1998a) relies on: If he can show that values of $P(C|R=k)$ fall below .5, he believes that this will refute the claim that the decision criterion is placed at the optimal position and will provide evidence that it remains at the unbiased position when payoffs or probabilities are varied, contrary to SDT.

In Figure 6B, the criterion is unbiased: Although $P(s) = .1$, in this illustration x_5 has remained at the position appropriate for $P(s) = .5$. Thus, for Category 6, $f_n(x)$ is approximately equal to $f_s(x)$, but the signal is presented 1 time in 10. Accordingly, $P(C|R=6)$ is approximately .1 for the "6" response, and for higher category values, $P(C|R=k)$ rises. Figure 6A is the result expected for these criterion positions if Balakrishnan is correct in his belief that the decision criterion remains rigidly at the unbiased position.

In Figure 6D, the decision criterion is intermediate between the unbiased criterion and the optimal criterion, giving $l(x) = 5$. For these parameters, $P(C|R=6)$ is approximately .35, intermediate between .1 and .5, as is shown in Figure 6C.

Figures 6A, 6C, and 6E demonstrate that the results for $P(C|R=k)$ provide indications of the extent of the underlying criterion shifts. In Figure 6, the values for the

least confident *yes* response, “6,” are particularly useful. Examination of the $P(C|R = k)$ contours is not an ideal procedure, since the precise values and the shape of the contour are affected by the criterion spacing. It is inferior to deriving predictions and testing goodness of fit, but that requires that we make distributional assumptions, a procedure Balakrishnan rejects.

Balakrishnan (1998a) reports a number of results for $P(s) = .1$ in which $P(C|R = k)$ for low-confidence *yes* responses falls below .5. He argues that these findings are “consistent with the interpretation that the decision rule was in fact unbiased” (Balakrishnan, 1998a, p. 613).

There are two errors in the argument that leads Balakrishnan to believe these results provide further support for his thesis. The first is that he is concerned only whether $P(C|R = k)$ is greater than or less than .5 and fails to consider the quantitative implications of the actual values obtained. The second is that he considers only the two cases in which the decision criterion is either unbiased or else at the optimal position defined by SDT and fails to take into account that SDT studies have often found an incomplete shift of x_c toward the optimal position (Green & Swets, 1966).

The results in his Figures 5 and 7 show that the probabilities of a correct response, for the least confident *yes* response, range between about .25 and .45. In noting only that these values are less than .5 and thus consistent with a constant decision criterion, Balakrishnan (1998a) fails to observe that they are equally consistent with SDT: Figure 6C illustrates the sort of result that may be obtained when the criterion shifts in response to the asymmetry of the probabilities, but not to the full theoretical extent. In Figure 6C, $P(C|R = 6)$ is approximately .35. This value lies within the range he obtained (approximately .25–.45) for the least confident *yes* response.

Balakrishnan also fails to consider the quantitative implications of the values he obtains for the least confident *yes* response. If the decision criterion stayed rigidly at the location at which $l(x_c) = 1$, even when $P(s) = .1$, and if the categories neighboring it are narrow, the probability that the least confident *yes* response is correct should approach .1, as illustrated in Figure 6A. But his data do not give so low a value. Although the exact values to be expected depend on the criterion spacing, the results he obtains seem most consistent with a partial shift of the decision criterion toward the optimal position and, thus, evidence for rather than against SDT.

Discussion and Conclusions

Balakrishnan (1998b) proposed a new sensitivity index, S' , that he claims is “better justified and statistically more reliable than . . . d' ” (Balakrishnan, 1998b, p. 69). Examination of the index shows that this wide general claim cannot be sustained. Over an extensive range of SDT procedures and designs, d' has advantages over S' , or S' cannot be calculated at all. However, there is a subset of designs, discussed above, in which use of S' may have an advantage, and in drawing this to our attention it appears that Balakrishnan may have made a valuable contribution.

Balakrishnan (1998a, 1998b, 1999) introduced three further tests or measures of response bias that he claimed establish that the decision criterion stays rigidly at the value that gives $l(x_c) = 1$, despite changes in payoffs or probabilities, and thus shows that SDT is fundamentally flawed. Balakrishnan (1998a) considered his data to show “no effect on response bias as this notion is defined in signal detection theory” (p. 614). Similarly, Balakrishnan (1999) concluded that the decision rule is unbiased despite payoff and base rate manipulations: “This seemingly paradoxical result is fundamentally inconsistent with the entire family of signal detection theory models, raising some important questions about the significance of many published results” (p. 1189). These conclusions, if correct, would overthrow half a century of work in SDT.

However, careful examination of the arguments and procedures underlying the three tests has shown serious flaws in each. We conclude that the evidence he has produced is not of sufficient weight to justify his devastating conclusion. It is Balakrishnan’s arguments that fail, not SDT.

Finally, one may note that Balakrishnan (1999) offered an alternative to the SDT account in order to explain how it is that payoff and probability manipulations alter false positive rates even though, he claimed, biasing of the decision criterion does not occur. His proposal was that payoffs and probabilities affect the variances of the noise and signal density functions, without biasing the decision criterion; he referred to this as “encoding plasticity and decision-making rigidity.” Thus, if the probability of the signal is reduced, he assumes that the variance of the signal density function is increased and that of noise decreased. If the decision criterion is located at the intersection of the two density functions, the hit and false positive rates decrease.

There are two difficulties with this argument. The first is that SDT manipulations affect not only hit rates and false positive rates, but also the value of β . The variance adjustment model does not explain the last finding. Second, it is difficult to see how the stimulus-dependent changes in the two variances could be achieved. The random noise that determines the variances of the central effects of the signal and noise stimuli can be thought of as arising from three sources. First, there is physical variability in the production of the stimulus. Second, there is the variation in the response of neurons and in transmission across synapses along the pathway by which the sensory input is transmitted to the decision axis prior to the decision—so-called *neural* or *biological* noise. Third, the decision criterion may be subject to variance. The physical variability of the stimulus is not under the organism’s control, and the same criterion is employed in making decisions about all inputs. Thus, Balakrishnan’s account is reduced to the hypothesis that the organism can vary its level of neural noise differentially for the signal stimulus and the noise stimulus. When the probability of the signal is low, there would be increased neural noise during the transmission of the sensory message from the signal to the decision axis and decreased noise during transmission of the sensory message from the noise stimulus.

Thus, for example, for a block of 50 signal and 50 noise stimuli, randomly ordered, the perceptual system would perform one operation on the sensory pathways on each of the 50 trials on which the signal is presented and a different operation on the 50 trials on which noise is presented, each of these operations having its effect on the sensory message during its transmission to the decision axis, so that the variance of the central effects recorded on the decision axis for the 50 signal presentations would be greater than the corresponding variance for the noise trials. But how does the organism know whether an input in process of transmission to the decision axis comes from signal or noise? This information becomes available only once the sensory effect at the decision axis has been compared with the criterion. This is too late to trigger the operation that should have been performed during the prior transmission of that sensory message to the decision axis. Until an adequate answer to this question can be produced, Balakrishnan's proposed alternative to the SDT model will have little further claim on our attention.

REFERENCES

- BALAKRISHNAN, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, **40**, 601-623.
- BALAKRISHNAN, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, **3**, 68-90.
- BALAKRISHNAN, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 1189-1206.
- BALAKRISHNAN, J. D., & RATCLIFF, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 615-633.
- EGAN, J. P., GREENBERG, G. Z., & SCHULMAN, A. I. (1961). Interval of time uncertainty in auditory detection. *Journal of the Acoustical Society of America*, **33**, 771-778.
- 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. (1960). Psychoacoustics and detection theory. *Journal of the Acoustical Society of America*, **32**, 1189-1203.
- GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- McNICOL, D. (1972). *A primer of signal detection theory*. London: George Allen and Unwin.
- STRETCH, V., & WIXTED, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.
- SWETS, J. A. (Ed.) (1964). *Signal detection and recognition by human observers*. New York: Wiley.
- TANNER, W. P., & SWETS, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**, 401-409.
- TREISMAN, M. (1999). There are two types of psychometric function: A theory of cue combination in the processing of complex stimuli with implications for categorical perception. *Journal of Experimental Psychology: General*, **128**, 517-546.
- TREISMAN, M., & FAULKNER, A. (1984a). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical & Statistical Psychology*, **37**, 199-215.
- TREISMAN, M., & FAULKNER, A. (1984b). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 119-139.
- TREISMAN, M., & FAULKNER, A. (1985). Can decision criteria interchange locations? Some positive evidence. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 187-208.
- TREISMAN, M., & WATTS, T. R. (1966). Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds: Estimating d' from results given by the method of constant stimuli. *Psychological Bulletin*, **66**, 438-454.
- TREISMAN, M., & WILLIAMS, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, **91**, 68-111.

(Manuscript received February 29, 2000;
revision accepted for publication November 7, 2001.)